Scanning and Sequential Decision Making for Multi-Dimensional Data - Part II: the Noisy Case[*]

Asaf Cohen[†], Tsachy Weissman[‡]and Neri Merhav[†]

February 5, 2008

Abstract

We consider the problem of sequential decision making on random fields corrupted by noise. In this scenario, the decision maker observes a noisy version of the data, yet judged with respect to the clean data. In particular, we first consider the problem of sequentially scanning and filtering noisy random fields. In this case, the sequential filter is given the freedom to choose the path over which it traverses the random field (e.g., noisy image or video sequence), thus it is natural to ask what is the best achievable performance and how sensitive this performance is to the choice of the scan. We formally define the problem of scanning and filtering, derive a bound on the best achievable performance and quantify the excess loss occurring when non-optimal scanners are used, compared to optimal scanning and filtering.

We then discuss the problem of sequential scanning and prediction of noisy random fields. This setting is a natural model for applications such as restoration and coding of noisy images. We formally define the problem of scanning and prediction of a noisy multidimensional array and relate the optimal performance to the clean scandictability defined by Merhav and Weissman. Moreover, bounds on the excess loss due to sub-optimal scans are derived, and a universal prediction algorithm is suggested.

[*] The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Seattle, Washington, United States, July 2006, and accepted to the IEEE International Symposium on Information Theory, Nice, France, June 2007.

[†] Asaf Cohen and Neri Merhav are with the Department of the Electrical Engineering, Technion - I.I.T., Haifa 32000, Israel. E-mails: {soofsoof@tx,merhav@ee}.technion.ac.il.

[‡] Tsachy Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. E-mail: tsachy@stanford.edu.

This paper is the second part of a two-part paper. The first paper dealt with sequential decision making on noiseless data arrays, namely, when the decision maker is judged with respect to the same data array it observes.

1 Introduction

Consider the problem of sequentially scanning and filtering (or predicting) a multidimensional noisy data array, while minimizing a given loss function. Particularly, at each time instant t, $1 \leq t \leq |B|$, where $|B|$ is the number of sites ("pixels") in the data array, the sequential decision maker chooses a site to be visited, denoted by Ψ_t. In the filtering scenario, it first observes the value at that site, and then gives an estimation for the underlying clean value. In the prediction scenario, it is required to give a prediction for that (clean) value, before the actual observation is made. In both cases, both the location Ψ_t and the estimation or prediction may depend on the previously observed values - the values at sites Ψ_1 to Ψ_{t-1}. The goal is to minimize the cumulative loss after scanning the entire data array.

Applications of this problem can be found in image and video processing, such as filtering or predictive coding. In these applications, one wishes to either enhance or jointly enhance and code a given image. The motivation behind a prediction/compression-based approach, is that the prediction error may consist mainly of the noise signal, while the clean signal is recovered by the predictor. For example, see [1]. It is clear that different scanning patterns of the image may result in different filtering or prediction errors, thus, it is natural to ask what is the performance of the optimal scanning strategy, and what is the loss when non-optimal strategies are used.

The problem of scanning multidimensional data arrays also arises in other areas of image processing, such as one-dimensional wavelet [2] or median [3] processing of images, where one seeks a space-filling curve which facilitates the one-dimensional signal processing of the multidimensional data. Other examples include digital halftoning [4], where a space filling curve is sought in order to minimize the effect of false contours, and pattern recognition [5]. Yet more applications can be found in multidimensional data query [6] and indexing [7], where multidimensional data is stored on a one-dimensional storage device, hence a locality-preserving space-filling curve is sought in order to minimize the number of continuous read operations required to access a multidimensional object, and rendering of three-dimensional graphics [8], [9].

An information theoretic discussion of the scanning problem was initiated by Lempel and Ziv in [10], where the Peano-Hilbert scan was shown to be optimal for compression of individual images. In [11], Merhav and Weissman formally defined a "scandictor", a scheme for sequentially scanning and predicting a multidimensional data array, as well as the "scandictability" of a random field, namely, the best achievable performance for scanning and prediction of a random field. Particular cases where this value can be computed and the optimal scanning order can be identified were discussed in that work. One of the main results of [11] is the fact that if a stochastic field can be represented autoregressively (under a specific scan Ψ) with a maximum-entropy innovation process, then it is optimally scandicted in the way it was created (i.e., by the specific scan Ψ and its corresponding optimal predictor). A more comprehensive survey can be found in [12] and [13]. In [12], the problem of *universal* scanning and prediction of noise-free multidimensional arrays was investigated. Although this problem is fundamentally different from its one-dimensional analogue (for example, one cannot compete successfully with any two scandictors on any individual image), a universal scanning and prediction algorithm which achieves the scandictability of any stationary random field was given, and the excess loss incurred when non-optimal scanning strategies are used was quantified.

In [14], Weissman, Merhav and Somekh-Baruch, as well as Weissman and Merhav in [15] and [16], extended the problem of universal prediction to the case of a *noisy* environment. Namely, the predictor observes a noisy version of the sequence, yet, it is judged with respect to the clean sequence. In this paper, we extend the results of [11] and [12] to this noisy scenario. We formally define the problem of sequentially filtering or predicting a multidimensional data array. First, we derive lower bounds on the best achievable performance. We then discuss the scenario where non-optimal scanning strategies are used. That is, we assume that, due to implementation constraints, for example, one cannot use the optimal scanner for a given data array, and is forced to use an arbitrary scanning order. In such a scenario, it is important to understand what is the excess loss incurred, compared to optimal scanning and filtering (or prediction). We derive upper bounds on this excess loss. Finally, we briefly mention how the results of [12] can be exploited in order to construct universal schemes to the noisy case as well. While many of the results for noisy scandiction are extendible from the noiseless case, similarly as results for noisy prediction were extended from results for noiseless prediction [15], the scanning and filtering problem poses new challenges and requires the use of new tools and techniques.

The paper is organized as follows. Section 2 includes a precise formulation of the problem.

Section 3 includes the results on scanning and filtering of noisy data arrays, while Section 4 is devoted to the prediction scenario. In both sections, particular emphasis is given to the important cases of Gaussian random fields corrupted by Additive White Gaussian Noise (AWGN), under the squared error criterion, and binary random fields corrupted by a Binary Symmetric Channel (BSC), under the Hamming loss criterion.

In particular, in Section 3.1, a new tool is used to derive a lower bound on the optimum scanning and filtering performance (Section 4.1 later shows how this tool can be used to strengthen the results of [11] in the noise-free scenario as well). Section 3.2 gives upper bounds on the excess loss in non-optimal scanning. In Section 3.2.1, the results of Duncan [17] as well as those of Guo, Shamai and Verdú [18] are used to derive the bounds when the noise is Gaussian, and Section 3.2.2 deals with the binary setting. Section 3.3 uses recent results by Weissman *et. al.* [19] to describe how universal scanning and filtering algorithms can be constructed. In the noisy scandiction section, Section 4.1 relates the best achievable performance in this setting, as well as the achieving scandictors, to the *clean scandictability* of the noisy field. Section 4.2 introduces a universal scandiction algorithm, and Section 4.3 gives an upper bound on the excess loss. In both Section 3 and Section 4, the sub-sections describing the optimum performance, the excess loss bounds and the universal algorithms are not directly related and can be read independently. Finally, Section 5 contains some concluding remarks.

2 Problem Formulation

We start with a formal definition of the problem. Let A denote the alphabet, which is either discrete or the real line. Let N be the noisy observation alphabet. Let $\Omega = (A \times N)^{\mathbb{Z}^2}$ be the observation space (the results can be extended to any finite dimension). A probability measure Q on Ω is stationary if it is invariant under translations τ_i, where for each $\omega \in \Omega$ and $i, j \in \mathbb{Z}^2$, $\tau_i(\omega)_j = \omega_{j+i}$ (namely, stationarity means shift invariance). Denote by $\mathcal{M}(\Omega)$ and $\mathcal{M}_S(\Omega)$ the sets of all probability measures on Ω and stationary probability measures on Ω, respectively. Elements of $\mathcal{M}(\Omega)$, *random fields*, will be denoted by upper case letters while elements of Ω, *individual data arrays*, will be denoted by the corresponding lower case. It will also be beneficial to refer to the clean and noisy random fields separately, that is, $\{X_t\}_{t\in\mathbb{Z}^2}$ represents the clean signal and $\{Y_t\}_{t\in\mathbb{Z}^2}$ represents the noisy observations, where for $t \in \mathbb{Z}^2$, X_t is the random variable corresponding to X at site t.

Let $\mathcal{V}$ denote the set of all finite subsets of $\mathbb{Z}^2$. For $V \in \mathcal{V}$, denote by X_V the restrictions of the data array X to V. Let $\mathcal{R}_\Box$ be the set of all rectangles of the form $V = \mathbb{Z}^2 \cap ([m_1, m_2] \times [n_1, n_2])$. As a special case, denote by V_n the square $\{0, \ldots, n-1\} \times \{0, \ldots, n-1\}$. For $V \subset \mathbb{Z}^2$, let the interior radius of V be

$$R(V) \stackrel{\triangle}{=} \sup\{r : \exists c \text{ s.t. } B(c,r) \subseteq V\}, \tag{1}$$

where $B(c,r)$ is a closed ball (under the l_1-norm) of radius r centered at c. Throughout, $\ln(\cdot)$ will denote the natural logarithm.

Definition 1. A *scanner-filter pair* for a finite set of sites $B \in \mathcal{V}$ is the following pair (Ψ, F):

- The scan $\{\Psi_t\}_{t=1}^{|B|}$ is a sequence of measurable mappings, $\Psi_t : N^{t-1} \mapsto B$ determining the site to be visited at time t, with the property that

$$\left\{\Psi_1, \Psi_2(y_{\Psi_1}), \Psi_3(y_{\Psi_1}, y_{\Psi_2}), \ldots, \Psi_{|B|}\left(y_{\Psi_1}, \ldots, y_{\Psi_{|B|-1}}\right)\right\} = B, \quad \forall y \in N^B. \tag{2}$$

- $\{\tilde{F}_t\}_{t=1}^{|B|}$ is a sequence of measurable filters, where for each t, $\tilde{F}_t : N^t \mapsto D$ determines the reconstruction for the value at the site visited at time t, based on the current and previous observations, and D is the reconstruction alphabet.

Note that both the scanner Ψ and the filters $\{\tilde{F}_t\}$ base their decisions only on the noisy observations. In the prediction scenario (i.e., noisy scandiction), we define $F_t : N^{t-1} \mapsto D$, that is, $\{F_t\}$ represents measurable predictors, which have access only to previous observations. We allow *randomized* scanner-filter pairs, namely, pairs such that $\{\Psi_t\}_{t=1}^{|B|}$ or $\{\tilde{F}_t\}_{t=1}^{|B|}$ can be chosen randomly from some set of possible functions. It is also important to note that we consider only scanners for finite sets of sites, ones which can be viewed merely as a reordering of the sites in a finite set B.

The cumulative loss of a scanner-filter pair $(\Psi, \tilde{F})$ up to time $t \leq |B|$ is denoted by $L_{(\Psi,\tilde{F})}(x_B, y_B)_t$,

$$L_{(\Psi,\tilde{F})}(x_B, y_B)_t = \sum_{i=1}^{t} l(x_{\Psi_i}, \tilde{F}_i(y_{\Psi_1}, \ldots, y_{\Psi_i})), \tag{3}$$

where $l : A \times D \mapsto [0, \infty)$ is the loss function. The sum of the instantaneous losses over the entire data array B, $L_{(\Psi,\tilde{F})}(x_B, y_B)_{|B|}$, will be abbreviated as $L_{(\Psi,\tilde{F})}(x_B, y_B)$.

For a given loss function l and a field $Q \in \mathcal{M}(\Omega)$ restricted to B, define the best achievable scanning and filtering performance by

$$\tilde{U}(l, Q_B) = \inf_{(\Psi,\tilde{F}) \in \mathcal{S}(B)} E_{Q_B} \frac{1}{|B|} L_{(\Psi,\tilde{F})}(X_B, Y_B), \tag{4}$$

where Q_B is the marginal probability measure restricted to B and $\mathcal{S}(B)$ is the set of *all* possible scanner-filter pairs for B. The best achievable performance for the field Q, $\tilde{U}(l, Q)$, is defined by

$$\tilde{U}(l, Q) = \lim_{n \to \infty} \tilde{U}(l, Q_{V_n}), \tag{5}$$

if this limit exists.

In the prediction scenario, F_t is allowed to base its estimation only on $y_{\Psi_1}, \ldots, y_{\Psi_{t-1}}$, and we have

$$L_{(\Psi,F)}(x_B, y_B) = \sum_{t=1}^{|B|} l(x_{\Psi_t}, F_t(y_{\Psi_1}, \ldots, y_{\Psi_{t-1}})), \tag{6}$$

$$\bar{U}(l, Q_B) = \inf_{(\Psi,F)} E_{Q_B} \frac{1}{|B|} L_{(\Psi,F)}(X_B, Y_B), \tag{7}$$

and

$$\bar{U}(l, Q) = \lim_{n \to \infty} \bar{U}(l, Q_{V_n}), \tag{8}$$

if this limit exists.

The following proposition asserts that for any stationary random field both the limit in (5) and the limit in (8) exist.

Proposition 1. *For any stationary field $Q \in \mathcal{M}_S(\Omega)$ and for any sequence $\{B_n\}$, $B_n \in \mathcal{R}_\square$, satisfying $R(B_n) \to \infty$, the limits in* (5) *and* (8) *exist and satisfy*

$$\tilde{U}(l, Q) = \lim_{n \to \infty} \tilde{U}(l, Q_{B_n}) = \inf_{\Delta \in \mathcal{R}_\square} \tilde{U}(l, Q_\Delta), \tag{9}$$

$$\bar{U}(l, Q) = \lim_{n \to \infty} \bar{U}(l, Q_{B_n}) = \inf_{\Delta \in \mathcal{R}_\square} \bar{U}(l, Q_\Delta). \tag{10}$$

Since $\tilde{U}(l, Q_B)$ and $\bar{U}(l, Q_B)$, possess the sub-additivity property, e.g., for any $V, V', V \cap V' = \emptyset$, there exists a scanner-filter pair $(\Psi, \tilde{F})$ (or a scandictor (Ψ, F)) on $V \cup V'$ such that

$$E_Q L_{(\Psi,\tilde{F})}(X_{V \cup V'}, Y_{V \cup V'}) \leq |V| \tilde{U}(l, Q_V) + |V'| \tilde{U}(l, Q_{V'}), \tag{11}$$

the proof of Proposition 1 follows verbatim that of [11, Theorem 1].

3 Filtering of Noisy Data Arrays

In this section, we consider the scenario of scanning and filtering. In this case, a lower bound on the best achievable performance is derived. For the cases of Gaussian random fields corrupted by AWGN and binary valued fields observed through a BSC, we derive bounds on the excess loss when a non-optimal scanner is used (with an optimal filter). Finally, we briefly discuss universal scanning and filtering.

3.1 A Lower Bound on the Best Achievable Scanning and Filtering Performance

We assume an invertible memoryless channel, meaning the channel input distribution of a single symbol is uniquely determined given the output distribution. As an example, a discrete memoryless channel with an invertible channel matrix can be kept in mind. See [20] for a discussion on the conditions on the channel matrix for the invertibility property to hold. Moreover, as will be elaborated on later, the result below applies to more general channels, including continuous ones.

In the case of an invertible channel, we define associated Bayes envelope by

$$f_l(P) = \min_{g(\cdot)} El(X, g(Y)), \tag{12}$$

where P is the distribution of the channel output Y. Define

$$\zeta(d) = \max\{H(P) : f_l(P) \leq d\}, \tag{13}$$

and let $\bar{\zeta}(\cdot)$ be the upper concave ($\cap$) envelope of $\zeta(\cdot)$.

Theorem 2. *Let Y_B be the output of an invertible memoryless channel whose input is X_B. Then, for any scanner-filter pair $(\Psi, \tilde{F})$ we have*

$$\bar{\zeta}\left(\frac{1}{|B|} E_{Q_B} L_{(\Psi,\tilde{F})}(X_B, Y_B)\right) \geq \frac{1}{|B|} H(Y_B), \tag{14}$$

that is,

$$\bar{\zeta}\left(\tilde{U}(l, Q_B)\right) \geq \frac{1}{|B|} H(Y_B). \tag{15}$$

Proof. We prove the above theorem for the discrete case. Yet, the derivations below apply to the continuous case as well, with summations replaced by the appropriate integrals and the entropy replaced by differential entropy.

Denote by $\Psi(Y_B)$ the reordered output sequence, that is, $\{Y_{\Psi_1}, Y_{\Psi_2}, \ldots, Y_{\Psi_{|B|}}\}$. We have,

$$\begin{aligned}
H(Y_B) &\overset{(a)}{=} H(\Psi(Y_B)) \\
&= \sum_{t=1}^{|B|} H(Y_{\Psi_t}|Y^{\Psi_{t-1}}) \\
&= \sum_{t=1}^{|B|} \sum_{y^{\Psi_{t-1}}} H(Y_{\Psi_t}|Y^{\Psi_{t-1}} = y^{\Psi_{t-1}}) P(y^{\Psi_{t-1}}) \\
&\overset{(b)}{\leq} \sum_{t=1}^{|B|} \sum_{y^{\Psi_{t-1}}} \zeta \left(E_{Q_B} \left\{ l \left(X_{\Psi_t}, \tilde{F}_t(y^{\Psi_{t-1}}, Y_{\Psi_t}) \right) |Y^{\Psi_{t-1}} = y^{\Psi_{t-1}} \right\} \right) P(y^{\Psi_{t-1}}) \\
&\overset{(c)}{\leq} \sum_{t=1}^{|B|} \sum_{y^{\Psi_{t-1}}} \bar{\zeta} \left(E_{Q_B} \left\{ l \left(X_{\Psi_t}, \tilde{F}_t(y^{\Psi_{t-1}}, Y_{\Psi_t}) \right) |Y^{\Psi_{t-1}} = y^{\Psi_{t-1}} \right\} \right) P(y^{\Psi_{t-1}}) \\
&\overset{(d)}{\leq} \sum_{t=1}^{|B|} \bar{\zeta} \left(E_{Q_B} l \left(X_{\Psi_t}, \tilde{F}_t(Y^{\Psi_t}) \right) \right) \\
&\overset{(e)}{\leq} |B| \bar{\zeta} \left(\frac{1}{|B|} E_{Q_B} L_{\tilde{F}} \left(\Psi(X_B), \Psi(Y_B) \right) \right) \\
&= |B| \bar{\zeta} \left(\frac{1}{|B|} E_{Q_B} L_{(\Psi, \tilde{F})}(X_B, Y_B) \right). \qquad (16)
\end{aligned}$$

The equality (a) is since the reordering does not change the entropy of Y_B. While this is clear for data-independent reordering, more caution is required when Ψ is a data-dependent scan. Yet, this can be proved using the chain rule, and noting that conditioned on $Y_{\Psi_1}^{\Psi_{t-1}}$, the next site Ψ_t is fixed (this is similar to the proof of [12, Proposition 13]). The inequalities (b) and (c) follow from the definitions of ζ and $\bar{\zeta}$ respectively, and (d) and (e) follow from Jensen's inequality. □

At this point, a few remarks are in order. Theorem 2 is the direct analogue of the lower bounds in [11] for the filtering scenario. Note, however, that it holds for any finite set of sites B. Furthermore, it applies to arbitrarily distributed random fields (even non-stationary fields), and to a wide family of loss functions. In fact, the only condition on $l(\cdot, \cdot)$ is that the associated Bayes envelope $f_l(P)$ is well defined. Note also that the lower bound on $\tilde{U}(l, Q)$ given in Theorem 2 results from the application of a *single letter* function, $\bar{\zeta}^{-1}(\cdot)$, to the normalized entropy of the noisy field, $\frac{1}{|B|}H(Y_B)$. That is, the memory in (X_B, Y_B) is reflected only in $\frac{1}{|B|}H(Y_B)$.

The proof of Theorem 2 is general and direct, however, it lacks the insightful geometrical interpretation which led to the lower bound in [11]. Therein, Merhav and Weissman showed that the transformation from a data array to an error sequence (defined by a specific scandictor

(Ψ, F)) is volume preserving. Thus, the least expected cumulative error is the *radius of a sphere*, whose volume is the volume of the set of all typical data arrays of the source. This happens when all the typical data arrays of the source map to a sphere in the "error vectors" space, and thus Merhav and Weissman were able to identify cases where the lower bound is tight. Currently, we cannot point out specific cases in which (15) is tight. Moreover, as the next two examples show, in the scanning and filtering scenario (unlike the scanning and prediction scenario we discuss in Section 4), $\zeta(d)$ may not be concave, and thus $\zeta(d) \neq \bar{\zeta}(d)$. Note, in this context, that there is no natural *time sharing* solution in this case, as there is no natural trade-off between two (or more) optimal points, and there is only one criterion to be minimized - the cumulative scanning and filtering loss (as opposed to rate versus distortion, for example).

3.1.1 Binary Input and BSC

To illustrate its use, we specialize Theorem 2 to the case of binary input through a BSC, i.e., the input random field X_{V_n} is binary, and Y_{V_n} is the output of a BSC whose input is X_{V_n} and crossover probability is $\delta < 1/2$. Note, however, that although the derivations below are specific for binary alphabet and Hamming loss, they are easily extendible to arbitrary finite alphabet and discrete memoryless channel with a channel transition matrix Π and loss function $\Lambda(\cdot, \cdot)$.

To compute the lower bound on the best achievable scanning and filtering performance, we evaluate $f_l(P)$ and $\zeta(d)$. By the definitions in (12) and (13), we consider the scalar problem of estimation of a random variable X based on its noisy observation Y. Denote by p_Y the probability $P(Y = 1)$ and by p_X the probability $P(X = 1)$. The best achievable performance,

$f_{l_H}(p_Y)$, which clearly depends on δ, and, hence, denoted $f_\delta(p_Y)$, is given by

$$\begin{aligned} f_\delta(p_Y) &= \sum_{x,y} P(x,y) l_H\left(x, g_{opt}(y)\right) \\ &= \sum_y P(y) \sum_x P(x|y) l_H\left(x, g_{opt}(y)\right) \\ &\overset{(a)}{=} \sum_y P(y) \min_x P(x|y) \\ &= \sum_y \min_x P(x,y) \\ &= \min\left\{p_X(1-\delta), \delta(1-p_X)\right\} + \min\left\{p_X\delta, (1-\delta)(1-p_X)\right\} \\ &= \min\{p_X, 1-p_X, \delta\} \\ &\overset{(b)}{=} \min\left\{\frac{p_Y-\delta}{1-2\delta}, \frac{1-p_Y-\delta}{1-2\delta}, \delta\right\}, \end{aligned} \tag{17}$$

where (a) results from the optimality of $g_{opt}(y)$ and (b) results from the invertability of the channel. Consequently,

$$\begin{aligned} \zeta(d) &= \max_p h_b(p) \quad \text{s.t.} \quad f_\delta(p) \le d \\ &= \begin{cases} h_b(\delta * d) & d < \delta \\ 1 & d \ge \delta, \end{cases} \end{aligned} \tag{18}$$

where $h_b(\cdot)$ is the binary entropy function and $\delta * d = d(1-\delta) + \delta(1-d)$. Note that since $\delta * \delta < 1/2$ for $0 < \delta < 1/2$, there is a discontinuity at $d = \delta$, hence $\zeta(d)$ is generally not concave and $\bar{\zeta}(d) \neq \zeta(d)$ (although $\bar{\zeta}(d)$ can be easily calculated). Figure 1 includes plots of both $\zeta(d)$ and $\zeta(\bar{d})$ for $\delta = 0.25$. We also mention that $d = \delta$ is a realistic cumulative loss in non-trivial situations, as there are cases where "say-what-you-see" (and thus suffer a loss δ) is the best any filter can do [21]. Furthermore, note that $\zeta(d)$ is not the maximum entropy function $\gamma(d)$ used in [11] to derive the lower bound on the scandictability.

Finally, exact evaluation of the bound in Theorem 2 may be difficult in many cases, as the entropy $\frac{1}{|B|}H(Y_B)$ may be hard to calculate, and only bounds on its value can be used.[1] At the end of Section 3.2.2, we give a numerical example for the bound in Theorem 2 using a lower bound on the entropy rate.

Remark 1. Clearly, $\zeta(d)$ is interesting only in the region $d \le \delta$, as any reasonable filter will have an expected normalized cumulative loss smaller or equal to the channel crossover probability. However, due to the discontinuity at $d = \delta$, $\zeta(d)$ is concave for $d < \delta$ but not for $d \le \delta$. This

[1] Think, for example, of an input process which is a first order Markov source. While the entropy rate of the input is known, the output is a hidden Markov process whose entropy rate is unknown in general.

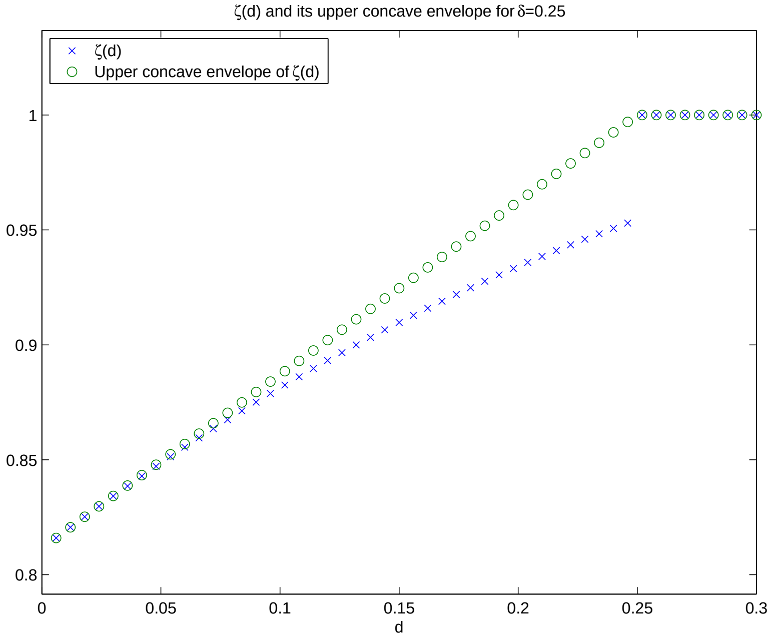

Figure 1: The function $\zeta(d)$, as it appears in (18), and its upper concave envelope, $\zeta(\bar{d})$, both plotted for $\delta = 0.25$. Note that $\zeta(d)$ and $\bar{\zeta}(d)$ have analytic expressions, and the plots are discrete only to better distinguish between them.

is fortunate, as if $\zeta(d)$ was concave on $d \leq \delta$, Theorem 2 would have resulted in $h_b(\delta * \delta)$ as an upper bound on the entropy rate of any binary source corrupted by a BSC, which is erroneous (for example, it violates $h_b(\pi * \delta)$ as a lower bound on the entropy rate of a first order Markov source with transition probability π corrupted by a BSC with crossover probability δ).

3.1.2 Gaussian Channel

Consider now the case where Y_{V_n} is the output of an AWGN channel, whose input is arbitrarily distributed. Assume the squared error loss. As the optimal filter is clearly the conditional expectation, $\zeta(d)$ in this case is given by

$$\zeta(d) = \max \left\{ H(X+N) : \mathrm{Var}(X|X+N) \leq d \right\}, \quad N \sim \mathcal{N}(0, \sigma_N^2), \quad N \perp X. \tag{19}$$

Since $H(X+N|X) = H(N)$ is fixed, this is similar to the classical Gaussian channel capacity problem, only now the input constraint is $\mathrm{Var}(X|X+N) \leq d$, which generally depends on the *distribution* of X rather than solely on its variance, and hence is not necessarily achieved by Gaussian X.

When the input is also limited to be Gaussian, however, the optimization problem in (13) is trivial and $\zeta(d)$ can be easily calculated (note that in this case $\zeta(d)$ is valid only to bound the performance for scanning and filtering of Gaussian fields corrupted by AWGN). Since the distributions depend only on the variance (assuming zero expectation), we have $f_{l_s}(P) = f_{l_s}(\sigma_Y^2)$, and, in fact,

$$\begin{aligned} f_{l_s}(\sigma_Y^2) &= \frac{\sigma_N^2 \sigma_X^2}{\sigma_N^2 + \sigma_X^2} \\ &= \sigma_N^2 - \frac{\sigma_N^4}{\sigma_Y^2}. \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned} \zeta(d) &= \max \frac{1}{2} \ln(2\pi e \sigma_Y^2) \quad \text{s.t.} \quad f_{l_s}(\sigma_Y^2) \le d \\ &= \begin{cases} \frac{1}{2} \ln\left(2\pi e \frac{\sigma_N^4}{\sigma_N^2 - d}\right) & d < \sigma_N^2 \\ \infty & d \ge \sigma_N^2. \end{cases} \end{aligned} \tag{21}$$

Unlike the binary setting, here the cumulative loss, d, will be strictly smaller than σ_N^2 for any non-trivial setting and reasonable filter, as the error in symbol-by-symbol filtering is $\sigma_N^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} < \sigma_N^2$. Yet, $\zeta(d)$ is convex ($\cup$) for $d < \sigma_N^2$, and the chain of inequalities in (16) cannot be tight.

3.2 Bounds on the Excess Loss of Non-Optimal Scanners

Theorem 2 gives a lower bound on the optimum scanning and filtering performance. However, it is interesting to investigate what is the excess scanning and filtering loss when non-optimal scanners are used. Specifically, in this section we address the following question: Suppose that, for practical reasons for example, one uses a non-optimal scanner, accompanied with the optimal filter for that scan. How large is the excess loss incurred by this scheme with respect to optimal scanning and filtering?

We consider both the case of a Gaussian channel and squared error loss (with Gaussian or arbitrarily distributed input) and the case of a binary source passed through a BSC and Hamming loss. While the tools we use in order to construct such a bound for the binary case are similar to the ones used in [12], we develop a new set of tools and techniques for the Gaussian setting.

3.2.1 Gaussian Channel

We investigate the excess scanning and filtering loss when non-optimal scanners are used, for the case of arbitrarily distributed input corrupted by a Gaussian channel. We first focus attention on the case where the input is Gaussian as well, and then derive a new results for the more general setting.

Similarly as in [12], the bound is achieved by bounding the absolute difference between the scanning and filtering performance of any two scans, Ψ^1 and Ψ^2, assuming both use their optimal filters. This bound, however, results from a relation between the performance of *discrete time* filtering and *continuous time* filtering, together with the fundamental result of Duncan [17] on the relation between mutual information and causal minimal mean square error estimation in a Gaussian channel. Namely, we use the *mutual information in continuous time* as a scan invariant feature, and the actual value of the excess loss bound results from the difference between discrete and continuous time filtering problems, as will be made precise below.

From now on we assume the loss function is the squared error loss, $l_s(\cdot)$. We start with several definitions. Let X be a Gaussian random variable, $X \sim \mathcal{N}(0, \sigma_X^2)$. Consider the following two estimation problems:

- The scalar problem of estimating X based on $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma_N^2)$, independent of X.
- The continuous time problem of causally estimating $X_t \equiv X$, $t \in [0,1]$, based on Y_t, which is an AWGN-corrupted version of X_t, the Gaussian noise having a spectral density level of σ_N^2.

To bound the sensitivity of the scanning and filtering performance, it is beneficial to consider the difference between the estimation errors in the above two problems, that is,

$$\int_0^1 \mathrm{Var}(X_t|Y^t)\mathrm{d}t - \mathrm{Var}(X|Y), \tag{22}$$

where Y^t is the continuous time signal $\{Y_{t'}\}_{t'=0}^t$. Clearly, $\mathrm{Var}(X|Y) = \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}$. Since $\int_0^t Y_{t'}\mathrm{d}t'$ is a sufficient statistics in the estimation of $X_t \equiv X$, $\mathrm{Var}(X_t|Y^t)$ is equivalent to the squared error in estimating X based on $X + \tilde{N}$, $\tilde{N}$ being a Gaussian random variable, independent of X, with zero mean and variance σ_N^2/t. Thus,

$$\begin{aligned}\int_0^1 \mathrm{Var}(X_t|Y^t)\mathrm{d}t - \mathrm{Var}(X|Y) &= \int_0^1 \frac{\sigma_X^2(\sigma_N^2/t)}{\sigma_X^2 + (\sigma_N^2/t)}\mathrm{d}t - \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2} \\ &= \sigma_N^2 \ln\left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right) - \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2} \\ &= \sigma_N^2 f\left(\frac{\sigma_X^2}{\sigma_N^2}\right),\end{aligned} \tag{23}$$

where

$$f(x) = \ln(1+x) - \frac{x}{x+1}. \tag{24}$$

The following is the main result in this sub-section.

Theorem 3. *Let X_{V_n} be a Gaussian random field with a constant marginal distribution satisfying $Var(X_i) = \sigma_X^2 < \infty$ for all $i \in V_n$. Let $Y_i = X_i + N_i$, where N_{V_n} is a white Gaussian noise of variance σ_N^2, independent of X_{V_n}. Then, for any two scans Ψ^1 and Ψ^2, we have*

$$\frac{1}{n^2}\left| EL_{(\Psi^1,\tilde{F}^{opt})}(X_{V_n}, Y_{V_n}) - EL_{(\Psi^2,\tilde{F}^{opt})}(X_{V_n}, Y_{V_n})\right| \leq \sigma_N^2 f\left(\frac{\sigma_X^2}{\sigma_N^2}\right). \tag{25}$$

Theorem 3 bounds the absolute difference between the scanning and filtering performance of any two scanners, Ψ^1 and Ψ^2, assuming they use their optimal filters. Clearly, since the scanners are arbitrary, this result can also be interpreted as the difference in performance between any scan Ψ, and the best achievable performance, $\tilde{U}(l, Q_{V_n})$. Note that the bound value, $\sigma_N^2 f\left(\frac{\sigma_X^2}{\sigma_N^2}\right)$, is a single letter expression, which depends on the input field X_{V_n} and the noise N_{V_n} only through their variances. Namely, the bound does not depend on the *memory in* X_{V_n}.

Proof (Theorem 3). As mentioned earlier, the comparison between any two scans is made by bounding the normalized cumulative loss of *any scan* Ψ in terms of a *scan invariant* entity, which is the mutual information.

For simplicity, assume first that the scan Ψ is data-independent, namely, it is merely a reordering of the entries of Y_{V_n}. In this case, $\{X_{\Psi_i}\}_{i=1}^{n^2}$ is a discrete time Gaussian vector. We construct from it a *continuous time process*, $\{X_t^{(c)}\}_{t\in[0,n^2]}$, where for any $t \in [i-1, i)$, $X_t^{(c)} = X_{\Psi_i}$, $i \in \{1, 2, \ldots, n^2\}$. That is, $X_t^{(c)}$ is a piecewise constant process, whose constant values at intervals of length 1 correspond to the original values of the discrete time vector $\{X_{\Psi_i}\}$. Let $\{Y_{\Psi_i}\}$ and $\{Y_t^{(c)}\}$ be the AWGN-corrupted versions on $\{X_{\Psi_i}\}$ and $X_t^{(c)}$, namely, $Y_{\Psi_i} = X_{\Psi_i} + N_{\Psi_i}$ and $Y_t^{(c)}$ is constructed according to

$$\mathrm{d}Y_t^{(c)} = X_t^{(c)}\mathrm{d}t + \sigma_N \mathrm{d}W_t, \quad t \in [0, n^2], \tag{26}$$

where W_t is a standard Brownian motion. Observe that the white Gaussian noise, $\sigma_N \mathrm{d}W_t$, has a spectral density of level σ_N^2, similar to the variance of the discrete time noise N_{V_n}. Since we switch from discrete time to continuous time, it is important to note that the noise value in the two problems is equivalent. That is, if the discrete time field X_{V_n} is corrupted by noise of variance σ_N^2, then we wish the continuous time white noise to have a spectrum such that the integral over an interval of length 1, whose integrand is the continuous output $Y_t^{(c)}$ (and thus is a sufficient statistics in order to estimate the piecewise-continuous input X_t in this interval), will be a random variable which is exactly $X_{\Psi_i} + N_{\Psi_i}$, N_{Ψ_i} having a variance of σ_N^2.

We have,

$$
\begin{aligned}
&\frac{1}{n^2} E_{Q_{V_n}} L_{(\Psi, \tilde{F}^{opt})}\left(X_{V_n}, Y_{V_n}\right) \\
&\qquad = \frac{1}{n^2} \sum_{i=1}^{n^2} \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_i}\right) \\
&\qquad \overset{(a)}{=} \frac{1}{n^2} \sum_{i=1}^{n^2} \Bigg[\int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_{t'}^{(c)}\}_{t' \in [i-1, i-1+t]}\right) \mathrm{d}t \\
&\qquad\qquad - \sigma_N^2 f\left(\frac{\mathrm{Var}\big(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}\big)}{\sigma_N^2}\right)\Bigg] \\
&\qquad \overset{(b)}{\geq} \frac{1}{n^2} \sum_{i=1}^{n^2} \int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_{t'}^{(c)}\}_{t' \in [i-1, i-1+t]}\right) \mathrm{d}t - \sigma_N^2 f\left(\frac{\mathrm{Var}\big(X_1\big)}{\sigma_N^2}\right) \\
&\qquad \overset{(c)}{=} \frac{1}{n^2} 2\sigma_N^2 I\left(\{X_t^{(c)}\}_{t \in [0, n^2]}; \{Y_t^{(c)}\}_{t \in [0, n^2]}\right) - \sigma_N^2 f\left(\frac{\mathrm{Var}\big(X_1\big)}{\sigma_N^2}\right) \\
&\qquad = \frac{1}{n^2} 2\sigma_N^2 I\left(\{X_{\Psi_i}\}; \{Y_{\Psi_i}\}\right) - \sigma_N^2 f\left(\frac{\mathrm{Var}\big(X_1\big)}{\sigma_N^2}\right) \\
&\qquad \overset{(d)}{=} \frac{1}{n^2} 2\sigma_N^2 I\left(X_{V_n}; Y_{V_n}\right) - \sigma_N^2 f\left(\frac{\mathrm{Var}\big(X_1\big)}{\sigma_N^2}\right).
\end{aligned}
\tag{27}
$$

The equality (a) is from the application of (23) with $X = X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}$, i.e., with X_{Ψ_i} distributed conditioned on $Y_{\Psi_1}^{\Psi_{i-1}}$. Note conditioned on $Y_{\Psi_1}^{\Psi_{i-1}}$, X_{Ψ_i} is indeed Gaussian, and that (23) applies to *any* Gaussian X corrupted by Gaussian noise. The inequality (b) is since $\mathrm{Var}(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}) \leq \mathrm{Var}(X_1)$ and due to the increasing monotonicity of f, (c) is since the resulting integral from 0 to n^2 is simply the minimal mean square error in filtering $\{Y_t^{(c)}\}$ (as Y_{Ψ_i} is a sufficient statistics with respect to $\{Y_{t'}^{(c)}\}_{t' \in [i-1, i-1+t]}$), and the application of Duncan's result [17, Theorem 3]. Finally, (d) is since the mutual information is invariant to the reordering of the random variables. To complete the proof of Theorem 3, simply note that since $f(x)$ is non-negative for $x > 0$, by (a) above, the normalized cumulative loss can be upper bounded as well, that is,

$$
\frac{1}{n^2} E_{Q_{V_n}} L_{(\Psi, \tilde{F}^{opt})}\left(X_{V_n}, Y_{V_n}\right) \leq \frac{1}{n^2} \sum_{i=1}^{n^2} \int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_{t'}^{(c)}\}_{t' \in [i-1, i-1+t]}\right) \mathrm{d}t \tag{28}
$$

hence, similarly as in the chain of inequalities leading to (27),

$$
\frac{1}{n^2} E_{Q_{V_n}} L_{(\Psi, \tilde{F}_{opt})}\left(X_{V_n}, Y_{V_n}\right) \leq \frac{1}{n^2} 2\sigma_N^2 I\left(X_{V_n}; Y_{V_n}\right). \tag{29}
$$

In fact, equation (29) can be viewed as the scanning and filtering analogue of [18, eq. (156a)].

Now, if the scan Ψ is data-dependent, the above derivations apply, with the use of the smoothing property of conditional expectation. That is, conditioned on $Y_{\Psi_1}^{\Psi_{i-1}}$, the position Ψ_i is fixed (assuming deterministic scanners, though random scanning order can be tackled with a similar method), relation (a) in (27) holds since it holds conditioned on $Y_{\Psi_1}^{\Psi_{i-1}}$, and relation (c) holds as the mutual information is invariant under data-dependent reordering as well. This is very similar to the methods used in the proof of [12, Proposition 13], where it was shown that the entropy of a vector is invariant to data-dependent reordering. □

At this point, a few remarks are in order. A very simple bound, applicable to arbitrarily distributed fields and under squared error loss (yet interesting mainly in the Gaussian regime) results from noting that for *any* random variables X and $Y = X + N$,

$$0 \leq \mathrm{Var}(X|Y) \leq \sigma_N^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}. \tag{30}$$

Namely, simple symbol by symbol restoration results in a cumulative loss of at most $\sigma_N^2 \frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2}$, and we have,

$$\begin{aligned} \frac{1}{n} EL_{(\Psi,\tilde{F}^{opt})}\left(\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n\right) &= \frac{1}{n}\sum_i \mathrm{Var}(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_i}) \\ &\leq \frac{1}{n}\sum_i \mathrm{Var}(X_{\Psi_i}|Y_{\Psi_i}) \\ &= \sigma_N^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}. \end{aligned} \tag{31}$$

Thus, the excess loss in non-optimal scanning cannot be greater than that value, hence,

$$\frac{1}{n^2}\left| EL_{(\Psi^1,\tilde{F}^{opt})}\left(X_{V_n}, Y_{V_n}\right) - EL_{(\Psi^2,\tilde{F}^{opt})}\left(X_{V_n}, Y_{V_n}\right)\right| \leq \sigma_N^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}. \tag{32}$$

In the next sub-section, we derive a tighter bound than the bound in (32), applicable to arbitrarily distributed noise-free fields. However, since this bound may be harder to evaluate, it is interesting to discuss the properties of (32) as well.

Both the bound in Theorem 3 and the bound in (32) are in the form of $\mathrm{Var}(X_1)g(\mathrm{SNR})$, for some g, where $\mathrm{SNR} = \sigma_X^2/\sigma_N^2$. This means that any bound obtained for a certain SNR applies to all values of $\mathrm{Var}(X_1)$ by rescaling. The bound in Theorem 3 has the form $\mathrm{Var}(X_1)\frac{f(\mathrm{SNR})}{\mathrm{SNR}}$, where $f(\cdot)$ was defined in (24), and we have

$$\lim_{\mathrm{SNR}\to 0^+} \frac{f(\mathrm{SNR})}{\mathrm{SNR}} = \lim_{\mathrm{SNR}\to\infty} \frac{f(\mathrm{SNR})}{\mathrm{SNR}} = 0, \tag{33}$$

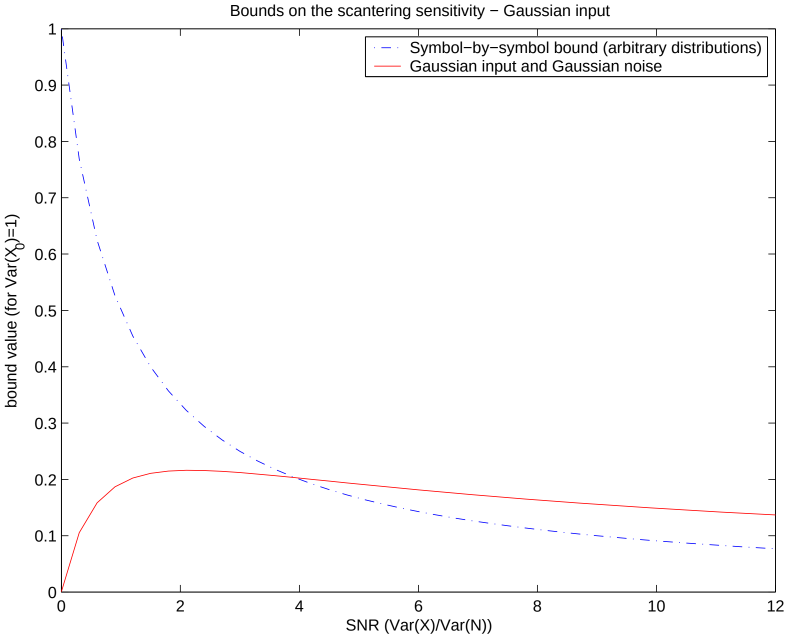

Figure 2: Bounds on the excess loss in scanning and filtering of Gaussian input corrupted by AWGN. The solid line is the bound given in Theorem 3. The dashed line is the bound given in (32).

that is, the scan is inconsequential at very high or very low SNR. This is clear as at high SNR the current observation is by far the most influential, and whatever previous observations used is inconsequential. For low SNR, the cumulative loss is high whatever the scan is. Unlike the bound in Theorem 3, (32) does not predict the correct behavior for SNR → 0^+, and is mainly interesting in the high SNR regime.

The above observations are also evident in Figure 2, which includes both the bound given in Theorem 3, applicable to Gaussian fields, and (32), applicable to arbitrarily distributed fields. It is also evident that in the case of Gaussian fields, $\frac{f(\mathrm{SNR})}{\mathrm{SNR}}$ has a unique maximum of approximately 0.216, that is, the excess loss due to a suboptimal scan at any SNR is upper bounded by $0.216\mathrm{Var}(X_1)$.

Remark 2. It is clear from the proof of Theorem 3 that an upper bound on the expression in (22), valid for arbitrarily distributed input X, may yield an upper bound on the excess scanning and filtering loss which is also valid for arbitrarily distributed random fields. However, while the integral in (22) can be upper bounded by assuming a Gaussian X, $\mathrm{Var}(X|Y)$ has no non-trivial lower bound. In fact, in [22], it is shown that if X is the following binary random variable,

$$X = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{w.p. } p, \\ -\sqrt{\frac{p}{1-p}} & \text{w.p. } 1-p, \end{cases} \tag{34}$$

for which $EX = 0$ and $EX^2 = 1$, then we have

$$Var(X|Y) \leq \frac{1}{2p(1-p)} e^{-\frac{\sigma_X^2/\sigma_N^2}{4p(1-p)}}, \tag{35}$$

which can be arbitrarily close to 0 for small enough p. Thus, the only lower bound on $V(X|Y)$ which is valid for any X with $EX^2 < \infty$, and depends only on σ_X^2 and σ_N^2, is 0 (and hence results is a bound weaker than Theorem 3 or (32)).

In the next two subsections, we derive new bounds on the excess loss, which are valid for more general input fields. First, we generalize the bound in Theorem 3. While the result may be complex to evaluate in its general form, we show that for binary input fields the bound admits a simple form. We then show that if the input alphabet is continuous, then a non-trivial bound on $\mathrm{Var}(X|Y)$ can be derived easily, which, in turn, results in a new bound on the excess loss.

A Generalization of Theorem 3. A generalization of Theorem 3 results from revisiting equality (a) of (27), which is simply the application of (23) with $X = X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{i-1}}$. While it is clear that an expression similar to that in (23) can be computed for non-Gaussian X, it is not clear that $X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{i-1}}$ has the same distribution for any $1 \leq i \leq n^2$ (unlike the Gaussian setting, where $X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{i-1}}$ is always Gaussian). Nevertheless, using the definition below, one can generalize Theorem 3 for arbitrarily distributed inputs as follows.

For any (X_{V_n}, Y_{V_n}), where Y_{V_n} is the AWGN-corrupted version of X_{V_n}, define

$$f^*(X_{V_n}, \sigma_N^2) = \max_{\Psi, 1 \leq i \leq n^2} \left\{ \int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_{t'}^{(c)}\}_{t' \in [i-1, i-1+t]} \right) \mathrm{d}t - \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_i} \right) \right\}. \tag{36}$$

Theorem 4. *Let X_{V_n} be an arbitrarily distributed random field, with a constant marginal distribution satisfying $Var(X_i) = \sigma_X^2 < \infty$ for all $i \in V_n$. Let $Y_i = X_i + N_i$, where N_{V_n} is a white Gaussian noise of variance σ_N^2, independent of X_{V_n}. Then, for any two scans Ψ^1 and Ψ^2, we have*

$$\frac{1}{n^2} \left| EL_{(\Psi^1, \tilde{F}^{opt})}(X_{V_n}, Y_{V_n}) - EL_{(\Psi^2, \tilde{F}^{opt})}(X_{V_n}, Y_{V_n}) \right| \leq f^*(X_{V_n}, \sigma_N^2). \tag{37}$$

The proof of Theorem 4 is similar to that of Theorem 3, and appears in Appendix A.1.

Note that $f^*(X_{V_n}, \sigma_N^2)$ is scan-independent, as it includes a maximization over all possible scans. At first sight, it seems like this maximization may take the sting out of the excess loss bound. However, as the example below shows, at least for the interesting scenario of binary input, this is not the case.

First, however, a few more general remarks are in order. Since important insight can be gained when using the results of Guo, Shamai and Verdú [18], let us mention the setting used therein. In [18], one wishes to estimate X based on $\sqrt{\mathrm{SNR}}X + N$, where N is a standard normal random variable. Denote by $\mathrm{I}(\mathrm{SNR})$ and $\mathrm{mmse}(\mathrm{SNR})$ the mutual information between X and $\sqrt{\mathrm{SNR}}X + N$, and the minimal mean square error in estimating X based on $\sqrt{\mathrm{SNR}}X + N$, respectively. Note that $\mathrm{Var}(X|Y)$ in our setting equals $\sigma_X^2 \mathrm{mmse}(\sigma_X^2/\sigma_N^2)$. Under these definitions,

$$\frac{\mathrm{d}}{\mathrm{dSNR}}\mathrm{I}(\mathrm{SNR}) = \frac{1}{2}\mathrm{mmse}(\mathrm{SNR}), \tag{38}$$

or, equivalently,

$$\mathrm{I}(\mathrm{SNR}) = \frac{1}{2}\int_0^{\mathrm{SNR}} \mathrm{mmse}(\gamma)\mathrm{d}\gamma. \tag{39}$$

Consequently, the result of Theorem 3 can be restated as

$$\frac{1}{n^2}\left|EL_{(\Psi^1,\tilde{F}^{opt})}(X_{V_n},Y_{V_n}) - EL_{(\Psi^2,\tilde{F}^{opt})}(X_{V_n},Y_{V_n})\right| \leq 2\sigma_N^2 \mathrm{I}(\mathrm{SNR}) - \sigma_X^2 \mathrm{mmse}(\mathrm{SNR}), \tag{40}$$

where $\mathrm{I}(\mathrm{SNR}) = \frac{1}{2}\ln(1+\mathrm{SNR})$ and $\mathrm{mmse}(\mathrm{SNR}) = \frac{1}{1+\mathrm{SNR}}$ are simply the mutual information and minimal mean square error of the *scalar* problem (hence, a single letter expression) of estimating a Gaussian X based on $\sqrt{\mathrm{SNR}}X + N$, where N standard Gaussian. In fact, the bound in Theorem 4 will always have the form $2\sigma_N^2 \mathrm{I}(\mathrm{SNR}) - \sigma_X^2 \mathrm{mmse}(\mathrm{SNR})$, for some X^* whose distribution is the maximizing distribution in (36). The next example shows that this is indeed the case for binary input as well, and the resulting bound can be easily computed.

Example 1 (*Binary input and AWGN*). Consider the case where X_{V_n} is a binary random field, with a symmetric marginal distribution (that is, $P(X_0 = \sigma_X) = P(X_0 = -\sigma_X) = 1/2$). Note that the X_i's are not necessarily i.i.d., and any dependence between them is possible. Y_{V_n} is the AWGN-corrupted version of X_{V_n}. To evaluate the bound in Theorem 4, $f^*(X_{V_n}, \sigma_N^2)$ should be calculated. However, for any scan Ψ and time i, $X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{i-1}}$ is still a binary random variable, taking the values $\pm\sigma_X$ with probabilities $(p, 1-p)$, for some $0 \leq p \leq 1/2$. Hence,

$$f^*(X_{V_n}, \sigma_N^2) \leq \max_{0 \leq p \leq 1/2}\left\{\int_0^1 \mathrm{Var}(X_t|Y^t)\mathrm{d}t - \mathrm{Var}(X|Y)\right\}, \tag{41}$$

where X is a binary random variable, taking the values $\pm\sigma_X$ with probabilities $(p, 1-p)$, $X_t \equiv X$, $Y = X + N$ and Y_t is the AWGN-corrupted version of X_t. The following result holds for any random variable X.

Claim 5. *For any random variable X with Var$(X) = \sigma_X^2 < \infty$, the expression in* (22) *is monotonically increasing in σ_X^2.*

Proof. We have,

$$
\begin{aligned}
\int_0^1 \mathrm{Var}(X|Y^t)\mathrm{d}t - \mathrm{Var}(X|Y) &= \int_0^1 \sigma_X^2 \mathrm{mmse}(\sigma_X^2/\sigma_N^2 t)\mathrm{d}t - \sigma_X^2 \mathrm{mmse}(\sigma_X^2/\sigma_N^2) \\
&= \sigma_N^2 \int_0^{\sigma_X^2/\sigma_N^2} \mathrm{mmse}(\gamma)\mathrm{d}\gamma - \sigma_X^2 \mathrm{mmse}(\sigma_X^2/\sigma_N^2). \qquad (42)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma_X^2}\left(\int_0^1 \mathrm{Var}(X_t|Y^t)\mathrm{d}t - \mathrm{Var}(X|Y)\right) &= -\sigma_X^2 \frac{\mathrm{d}}{\mathrm{d}\sigma_X^2}\mathrm{mmse}(\sigma_X^2/\sigma_N^2) \\
&= -2\mathrm{SNR}\frac{\mathrm{d}^2}{\mathrm{dSNR}^2}\mathrm{I}(\mathrm{SNR}) \\
&\geq 0, \qquad (43)
\end{aligned}
$$

where the last inequality is by [18, Corollary 1]. □

Claim 5 simply states that the monotonicity of $f(\cdot)$ used in inequality (b) of (27) is not specific for Gaussian input, and holds for any X. Thus, by Claim 5, the term in the braces of equation (41) is monotonically increasing in the variance of X, which is simply $4\sigma_X^2(p - p^2)$. Thus, it is maximized by $p = 1/2$, and we have

$$
f^*(X_{V_n}, \sigma_N^2) = 2\sigma_N^2 \mathrm{I}(\mathrm{SNR}) - \sigma_X^2 \mathrm{mmse}(\mathrm{SNR}), \qquad (44)
$$

where I(SNR) and mmse(SNR) are the mutual information and minimal mean square error in the estimation of X based on $\sqrt{\mathrm{SNR}}X + N$, where X is binary symmetric and N is a standard normal. Since the conditional mean estimate in this problem is $\tanh(\sqrt{\mathrm{SNR}}Y)$, we have [18]

$$
\mathrm{I}(\mathrm{SNR}) = \mathrm{SNR} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \ln\cosh(\mathrm{SNR} - \sqrt{\mathrm{SNR}}y)dy, \qquad (45)
$$

and

$$
\mathrm{mmse}(\mathrm{SNR}) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \tanh(\mathrm{SNR} - \sqrt{\mathrm{SNR}}y)dy, \qquad (46)
$$

so the bound can be computed numerically. The above bound is plotted in Figure 3. Similarly to the case of Gaussian input, it is insightful to compare this bound to a simple symbol-by-symbol filtering bound. That is, since for any binary X corrupted by AWGN of variance σ_N^2, $0 \leq \mathrm{Var}(X|Y) \leq \sigma_X^2 \mathrm{mmse}(\mathrm{SNR})$, we have

$$
\frac{1}{n^2}\left| EL_{(\Psi^1, \tilde{F}^{opt})}(X_{V_n}, Y_{V_n}) - EL_{(\Psi^2, \tilde{F}^{opt})}(X_{V_n}, Y_{V_n}) \right| \leq \sigma_X^2 \mathrm{mmse}(\mathrm{SNR}), \qquad (47)
$$

where mmse(SNR) is given in (46). This is simply the analogue of (32) to the binary input setting.

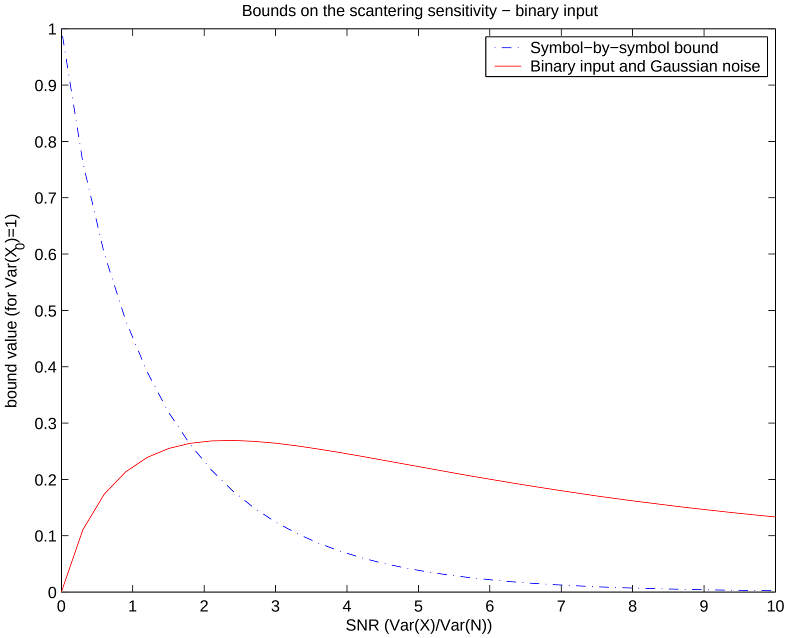

Figure 3: Bounds on the excess loss in scanning and filtering of binary input fields corrupted by AWGN. The solid line is the bound given in (44) (that is, Theorem 4), and the dashed line is the symbol-by-symbol bound given in (47).

A Bound for Arbitrarily Distributed Continuous Input. In this sub-section, we derive an additional bound on the excess scanning and filtering loss under squared error. We assume, however, that the input random field X_{V_n} is over $\mathbb{R}^{V_n}$, and that $X_i|Y_{V_n}$ has a finite differential entropy for any $i \in V_n$ (roughly speaking, this means that in the denoising problem of X_i, $X_i|Y_{V_n}$ is a non-degenerated continuous random variable). Under the above assumptions, we derive an excess loss bound which is not only valid for non-Gaussian input, but also depends on the memory in the random field (X_{V_n}, Y_{V_n}). On the other hand, it is important to note that the bound below is mainly asymptotic, and may be much harder to evaluate compared to the bounds in Theorem 3 or (32).

By [23, Theorem 9.6.5], for any X, Y with a finite conditional differential entropy $H(X|Y)$,

$$\mathrm{Var}(X|Y) \geq \frac{1}{2\pi e} \exp\left\{2H(X|Y)\right\}. \qquad (48)$$

Thus,

$$\begin{aligned}
\frac{1}{n}EL_{(\Psi,\tilde{F}^{opt})}(X_{V_n},Y_{V_n}) &= \frac{1}{n^2}\sum_{i=1}^{n^2}\mathrm{Var}(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_i}) \\
&\overset{(a)}{\geq} \frac{1}{n^2}\sum_{i=1}^{n^2}\frac{1}{2\pi e}\exp\{2H(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_i})\} \\
&\overset{(b)}{\geq} \frac{1}{n^2}\sum_{i=1}^{n^2}\frac{1}{2\pi e}\exp\{2H(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{n^2}})\} \\
&\overset{(c)}{\geq} \frac{1}{2\pi e}\exp\left\{2\frac{1}{n^2}\sum_{i=1}^{n^2}H(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{n^2}})\right\}, \qquad (49)
\end{aligned}$$

where (a) is by applying (48) with $Y = Y_{\Psi_1}^{\Psi_i}$, (b) is since conditioning reduces entropy and (c) is by applying Jensen's inequality.

The expression $\frac{1}{n^2}\sum_{i=1}^{n^2} H(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{n^2}})$ equals $\frac{1}{n^2}\sum_{i=1}^{n^2} H(X_{\Psi'_1}|Y_{\Psi'_1}^{\Psi'_{n^2}})$ for any two scanners Ψ and Ψ', since equality holds even without the expectation implicit in the entropy function. Thus, it is scan-invariant. Define

$$H^{+}(X|Y) = \liminf_{n\to\infty}\frac{1}{|V_n|}\sum_{i\in V_n}H(X_i|Y_{V_n}). \qquad (50)$$

$H^{+}(X|Y)$ can be seen as the asymptotic normalized entropy in the denoising problem of $\{X\}$ based on its noisy observations. Note that the entropies in (50) are differential. The following proposition gives a new lower bound on the excess scanning and filtering loss under squared error.

Proposition 6. *Let X_{V_n} be an arbitrarily distributed continuous valued random field with $Var(X_i) = \sigma_X^2$ for all i. Let $Y_i = X_i + N_i$, where N_{V_n} is a white noise of variance σ_N^2, independent of X_{V_n}. Assume that $X_i|Y_{V_n}$ has a finite differential entropy for any $i \in V_n$. Then, for any two scans Ψ^1 and Ψ^2, we have*

$$\begin{aligned}
\liminf_{n\to\infty}\frac{1}{|V_n|}&\left|EL_{(\Psi^1,\tilde{F}^{opt})}(X_{V_n},Y_{V_n}) - EL_{(\Psi^2,\tilde{F}^{opt})}(X_{V_n},Y_{V_n})\right| \\
&\leq \frac{\sigma_N^2\sigma_X^2}{\sigma_X^2+\sigma_N^2} - \frac{1}{2\pi e}\exp\left\{2H^{+}(X|Y)\right\}. \qquad (51)
\end{aligned}$$

Proof. The proof follows directly by applying the lower bound on the scanning and filtering performance given in (49) and the upper bound in (31). □

The bound in Proposition 6 is always at least as tight as the bound in (32) (and thus tighter than the bound in Theorem 3 for high SNR). For example, if the estimation error of X_i given

Y_{V_n} tends to zero as n increases (as in the case where $X_i = X$ for all i), then $\exp\{2H^+(X|Y)\} = 0$. However, if X_i cannot be reconstructed completely by Y_{V_n}, then the bound may be tighter than (32). It is far from being a tight bound on the excess loss, though. In the extreme case where all X_i's are i.i.d., the excess loss bound in Proposition 6 is $\frac{\sigma_N^2 \sigma_X^2}{\sigma_X^2 + \sigma_N^2} - \mathrm{Var}(X_1|Y_1) > 0$ (for non Gaussian X), while it is clear that all reasonable scanner-filter pairs perform the same. Finally, note that any lower bound on $H^+(X|Y)$ results in an upper bound on the scanning and filtering excess loss. For example, since

$$H^+(X|Y) \geq H(X_0|Y_{-k}^k, X_{-k-1}, X_{k+1}) \tag{52}$$

for any finite k, one can compute a simple upper bound on the excess loss, at least for first order Markov $\{X\}$.

3.2.2 Binary Input and BSC

Unlike the Gaussian setting discussed in Section 3.2.1, where the bound on the excess loss resulted from a continuous-time equality, with the *mutual information* serving as the scan-invariant feature, in the case of binary input and a BSC the *entropy* of the random field will play the key role, similar to [12]. As given in Section 3.1, the best achievable performance (in the scalar problem) is given by

$$f_\delta(p) = \min\left\{\frac{p-\delta}{1-2\delta}, \frac{1-p-\delta}{1-2\delta}, \delta\right\}, \tag{53}$$

where p is the probability that the channel output is 1 and δ is the channel crossover probability. Note that $f_\delta(p)$ is not the Bayes envelope associated with estimating X_t using Y_t under Hamming loss. However, as is clear from the derivations in (17), and will be evident from the proof of the following theorem, $f_\delta(P(y_t|y^{t-1}))$ is the *expectation* of the Bayes envelope (associated with estimating X_t using Y^t under Hamming loss) with respect to the distribution $P(y_t|y^{t-1})$. Define

$$\epsilon_\delta = \min_{a,b} \max_{\delta \leq p \leq 1/2} |a h_b(p) + b - f_\delta(p)|. \tag{54}$$

The following is the main result in this sub-section.

Theorem 7. *Let Y_B be the output of a BSC with crossover probability δ whose input is X_B. Then, for any scanner-filter pair $(\Psi, \tilde{F}^{opt})$, where $\tilde{F}^{opt}$ is the optimal filter for the scan Ψ, we have*

$$\left|\frac{1}{|B|} E_{Q_B} L_{\Psi, \tilde{F}^{opt}}(X_B, Y_B) - \tilde{U}(l_H, Q_B)\right| \leq 2\epsilon_\delta. \tag{55}$$

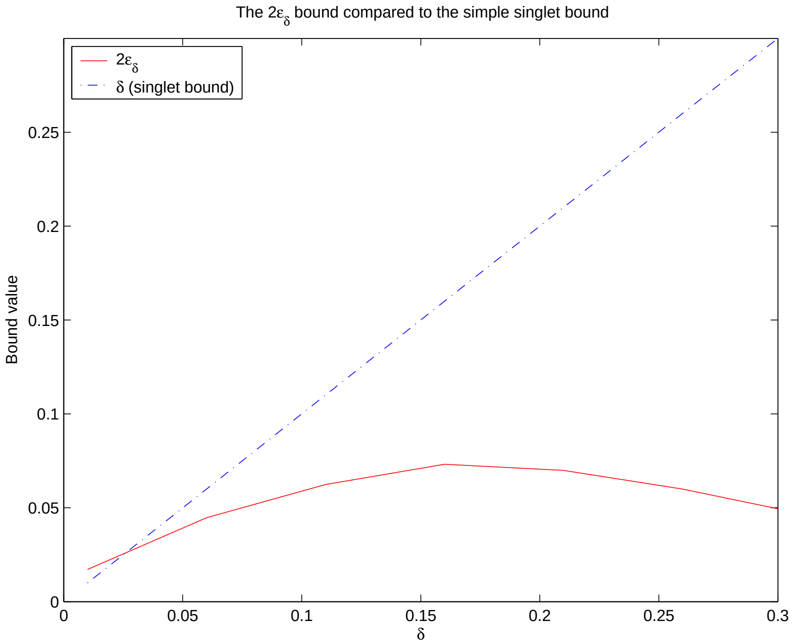

Figure 4: Bounds on the excess loss in scanning and filtering of binary random fields corrupted by a BSC. The solid line is the bound in Theorem 7 ($2\epsilon_\delta$), and the dashed line is the singlet bound ($f_\delta(1/2) = \delta$).

Even without evaluating ϵ_δ explicitly, it is easy to see that the excess loss when using non optimal scanners is quite small in this binary filtering scenario. For example, for $\delta = 0.1$ and $\delta = 0.25$ we have $\epsilon_\delta < 0.035$ and $\epsilon_\delta < 0.03$ respectively, yielding a maximal loss of 0.07 or even 0.06. Figure 4 includes the value of $2\epsilon_\delta$ as a function of δ. Similarly to Section 3.2.1, it is compared to a simple bound on the excess loss which results from simply bounding the Hamming loss of any filter by 0 from below and δ from above (namely, δ is the resulting bound on the excess loss). The values in Figure 4 should also be compared to 0.16, which is the bound on the excess loss in the clean *prediction* scenario [12], or even to larger values in the noisy prediction scenario, to be discussed in the next section. The fact that the filtering problem is less sensitive to the scanning order is quite clear as the noisy observation of X_{Ψ_t} is available under any scan. Finally, it is not hard to show that in the limits of $\delta \to 0$ and $\delta \to 1/2$ (high and low SNR, respectively), we have $\epsilon_\delta \to 0$, which is expected, as the scanning is inconsequential in these cases (note, however, that the singlet bound, δ, does not predict the correct behavior at low SNR).

Proof (Theorem 7). We first show that for any arbitrarily distributed binary n-tuple X^n and any $0 \leq \delta < 1/2$

$$\left| a_\delta^* \frac{1}{n} H(Y^n) + b_\delta^* - \frac{1}{n} EL_{l_H}^{opt}(X^n, Y^n) \right| \leq \epsilon_\delta, \tag{56}$$

where $EL_{l_H}^{opt}(X^n, Y^n)$ is the expected cumulative Hamming loss in optimally filtering X^n based on Y^n and a_δ^* and b_δ^* are the minimizers of ϵ_δ in (54). Indeed,

$$\begin{aligned}
&\left| a_\delta^* \frac{1}{n} H(Y^n) + b_\delta^* - \frac{1}{n} EL_{l_H}^{opt}(X^n, Y^n) \right| \\
&\quad = \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{y^{t-1}} P(y^{t-1}) \sum_{x_t, y_t} \Big[- a_\delta^* P(x_t, y_t | y^{t-1}) \ln P(y_t | y^{t-1}) + b_\delta^* P(x_t, y_t | y^{t-1}) \right. \\
&\qquad\qquad \left. - P(x_t, y_t | y^{t-1}) l_H \left(x_t, \tilde{F}^{opt}(y^t) \right) \Big] \right| \\
&\quad \leq \frac{1}{n} \sum_{t=1}^{n} \sum_{y^{t-1}} P(y^{t-1}) \left| a_\delta^* h_b \left(P(y_t | y^{t-1}) \right) + b_\delta^* - \sum_{x_t, y_t} P(x_t, y_t | y^{t-1}) l_H \left(x_t, \tilde{F}^{opt}(y^t) \right) \right| \\
&\quad \leq \frac{1}{n} \sum_{t=1}^{n} \sum_{y^{t-1}} P(y^{t-1}) \left| a_\delta^* h_b \left(P(y_t | y^{t-1}) \right) + b_\delta^* - \sum_{y_t} P(y_t | y^{t-1}) \sum_{x_t} P(x_t | y^t) l_H \left(x_t, \tilde{F}^{opt}(y^t) \right) \right|.
\end{aligned} \tag{57}$$

Consider the summation $\sum_{y_t} P(y_t|y^{t-1}) \sum_{x_t} P(x_t|y^t) l_H \left(x_t, \tilde{F}^{opt}(y^t) \right)$. As $\tilde{F}$ is optimal, the inner sum equals at most $\min\{P(x_t = 0|y^t), P(x_t = 1|y^t)\}$. Thus, similar to the derivations in (17), we have

$$\begin{aligned}
&\sum_{y_t} P(y_t|y^{t-1}) \sum_{x_t} P(x_t|y^t) l_H \left(x_t, \tilde{F}^{opt}(y^t) \right) \\
&\qquad = \sum_{y_t} P(y_t|y^{t-1}) \min\{P(x_t = 0|y^t), P(x_t = 1|y^t)\} \\
&\qquad = \min \left\{ \frac{P(y_t = 0|y^{t-1}) - \delta}{1 - 2\delta}, \frac{P(y_t = 1|y^{t-1}) - \delta}{1 - 2\delta}, \delta \right\}.
\end{aligned} \tag{58}$$

Let $p = P(y_t = 0|y^{t-1})$. Note that $\min\{p, 1-p\} \geq \delta$. We have

$$\begin{aligned}
&\left| a_\delta^* \frac{1}{n} H(Y^n) + b_\delta^* - \frac{1}{n} EL_{l_H}^{opt}(X^n, \delta) \right| \\
&\qquad \leq \frac{1}{n} \sum_{t=1}^{n} \sum_{y^{t-1}} P(y^{t-1}) \left| a_\delta^* h_b(p) + b_\delta^* - \min \left\{ \frac{p - \delta}{1 - 2\delta}, \frac{1 - p - \delta}{1 - 2\delta}, \delta \right\} \right| \\
&\qquad \leq \frac{1}{n} \sum_{t=1}^{n} \sum_{y^{t-1}} P(y^{t-1}) \max_{\delta \leq p \leq 1/2} |a_\delta^* h_b(p) + b_\delta^* - f_\delta(p)| \\
&\qquad = \epsilon_\delta,
\end{aligned} \tag{59}$$

which establishes (56). However, the same inequality can be proved for any reordering of the data Ψ (similar to the proof of [12, Proposition 13]), consequently,

$$\left| a_\delta^* \frac{1}{|B|} H(\Psi(Y_B)) + b_\delta^* - \frac{1}{|B|} E_{Q_B} L_{\Psi, \tilde{F}^{opt}}(X_B, Y_B) \right| \leq \epsilon_\delta. \tag{60}$$

Using (60), remembering that $H(Y_B) = H(\Psi(Y_B))$ for any Ψ, and applying the triangle inequality results in (55). $\square$

Note that analogous ideas were used by Verdú and Weissman to bound the absolute difference between the denoisability and erasure entropy [24].

Theorem 2 gives a lower bound on the best achievable scanning and filtering performance. Theorems 3, 4 and 7 give an upper bound on the maximal possible difference between the normalized cumulative loss of any two scanners (accompanied by the optimal filters), or any one scanner compared to the optimal scan. Although Theorem 2 is similar to the results of [11], even for the relatively simple examples of a Gaussian field through a Gaussian memoryless channel or a binary source through a binary symmetric channel we have no results which can parallel [11, Theorem 17] or [11, Corollary 21], i.e., give an example of an optimal scanner-filter pair for a certain scenario. However, as the next example shows, we can identify situations when scanning and filtering improves the filtering results, i.e., non trivial scanning of the data results in strictly better restoration. Moreover, the example below illustrates the use of the results derived in this section.

Example 2 (*One Dimensional Binary Markov Source and the BSC*). In this case, it is not too hard to construct a scheme in which non-trivial scanning improves the filtering performance. In [21], Ordentlich and Weissman study the optimality of symbol by symbol (singlet) filtering and decoding. That is, the regions (depending on the source and channel parameters) where a memoryless scheme to estimate X_i is optimal with respect to causal (filtering) or non causal (denoising) non-memoryless schemes. Clearly, in the regions where singlet denoising is optimal (a fortiori singlet filtering), scanning cannot improve the filtering performance. However, consider the region where singlet filtering is optimal, yet singlet denoising is not. In this region, there exists k for which the estimation error in estimating X_i based on Y_{i-k}^{i+k} is strictly smaller than that based on Y_i (as the optimal filter is memoryless yet the optimal denoiser is not). Hence, a scanner which in the first pass scans k contiguous symbols, then skips one, etc., and in the second pass returns to fill in the holes, accompanied by singlet filtering in the first pass and non-memoryless in the second, has strictly better filtering performance than the trivial scanner.

For a binary symmetric Markov source with a transition probability $\pi \leq \frac{1}{2}$, corrupted by a BSC with crossover probability δ, [21, Corollary 3] asserts that singlet filtering ("say-what-you-see" scheme in this case) is optimal if and only if

$$\delta \leq f(\pi) \stackrel{\triangle}{=} \frac{1}{2}(1 - \sqrt{\max\{1 - 4\pi, 0\}}). \tag{61}$$

Singlet denoising, on the other hand, is optimal if and only if

$$\delta \leq d(\pi) \stackrel{\triangle}{=} \frac{1}{2}\left(1 - \sqrt{\max\{1 - 4\left(\frac{\pi}{1-\pi}\right)^2, 0\}}\right). \tag{62}$$

Consider a scanner-filter pair which scans the data using an "odds-then-evens" scheme. On the odds, "say-what-you-see" filtering is used. On the evens, Y_{i-1}^{i+1} are used in order to estimate X_i.[2] The results are in Figures 5 and 6. In Figure 5, the points marked with "x" are where the "odds-then-evens" scan improves on the trivial scan. The two curves are $f(\pi)$ and $d(\pi)$. Figure 6 shows the actual improvement made by the "odds-then-evens" scanning and filtering.

For $\delta = \pi = 0.1$, for example, the "odds-then-evens" error rate is smaller than that of filtering with the trivial scan by 0.021 (that is, 0.079 compared to 0.1). This value should be put alongside the upper bound on the excess loss given in Theorem 7, which is smaller than 0.07 in this case. To evaluate the bound on the best achievable scanning and filtering performance given in Theorem 2 for this example (denoted, with a slight abuse of notation, as $\tilde{U}(\pi, \delta)$), we have

$$\begin{aligned} \tilde{U}(\pi, \delta) &\geq \bar{\zeta}^{-1}(\bar{H}(\pi, \delta)) \\ &\geq \bar{\zeta}^{-1}(h_b(\pi * \delta)), \end{aligned} \tag{63}$$

where $\bar{H}(\pi, \delta)$ is the entropy rate of the output, which is in turn lower bounded by $h_b(\pi * \delta)$. The resulting bound for $\pi = \delta = 0.1$ is approximately 0.04.

Thus, there exist non-trivial scanning and filtering schemes (i.e., lower bounds) whose improvement on the trivial scanning order is of the same order of magnitude as the upper bound in Theorem 7. To conclude, it is clear that there is a wide region were a non trivial scanning order improves on the trivial scan, an that this region includes at least all the region between $f(\pi)$ and $d(\pi)$. Yet, it is not clear what is the optimal scanner-filter pair.

3.3 Universal Scanning and Filtering of Noisy Data Arrays

In [19], Weissman *et. al.* mention that the problems involving sequential decision making on noisy data are not fundamentally different from their noiseless analogue, and in fact can be

[2] This is to have few simple steps in the forward-backward algorithm [25, Section 5] which is required to compute $P(x_i|y_{i-1}^{i+1})$. The generalization to Y_{i-k}^{i+k} is straightforward.

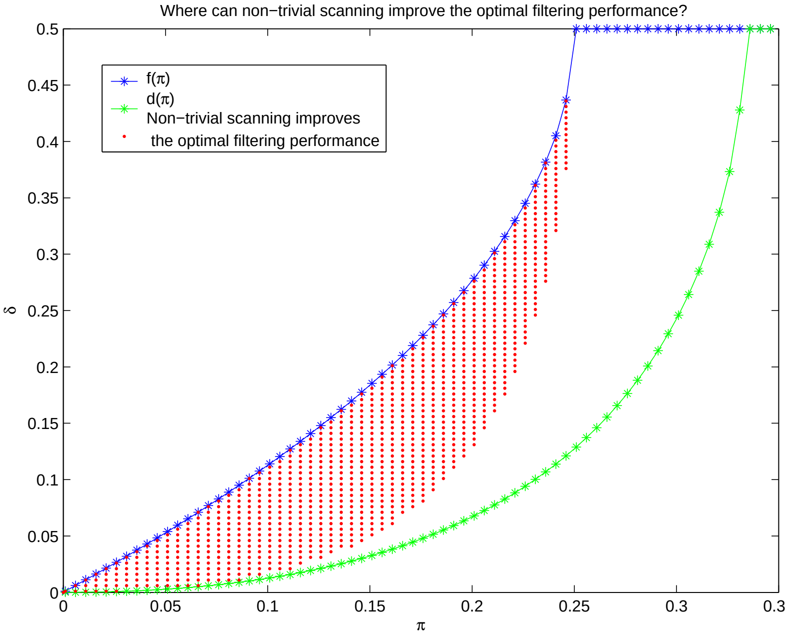

Figure 5: Where can a simple (suboptimal) "odds-then-evens" scan improve on the trivial scanning order and optimal filtering scheme. π is the transition probability of the symmetric, first order, Markov source and δ is the channel crossover probability.

reduced to the noiseless setting using a properly modified loss function. Indeed, this property of the noisy setting was used throughout the literature, and in this work. The problem of filtering a noisy data sequence is not different in this sense, and it is possible to construct a modified loss function such that the filtering problem is transformed into a prediction problem (with a few important exceptions to be discussed later). Such a modified loss function and a 'filtering-prediction transformation' is discussed in [19]. We briefly review this transformation, and consider its use in universal filtering of noisy data arrays.

First, we slightly generalize our notion of a filter. For a random variable U_t uniformly distributed on $[0, 1]$, let $\hat{X}_t(y^{t-1}y_t, U_t) \in A$ denote the output of the filter $\hat{X}$ at time t, after observing y^t. That is, the filter $\hat{X}$ also views an auxiliary random variable, on which it can base its output, $\hat{X}_t(y^{t-1}y_t, U_t)$. We also generalize the prediction space to $\mathcal{M}(\mathcal{S})$, $\mathcal{S} = \{s : N \mapsto A\}$. I.e., the prediction space is a distribution on the set of functions from the noisy observations alphabet N to the clean signal alphabet A. We assume an invertible discrete memoryless channel.

For each filter $\hat{X}$, the corresponding predictor is defined by

$$F_t^{\hat{X}}(y^{t-1})[s] = P\left(\hat{X}_t(y^{t-1}y, U_t) = s(y) \quad \forall y \in N\right). \tag{64}$$

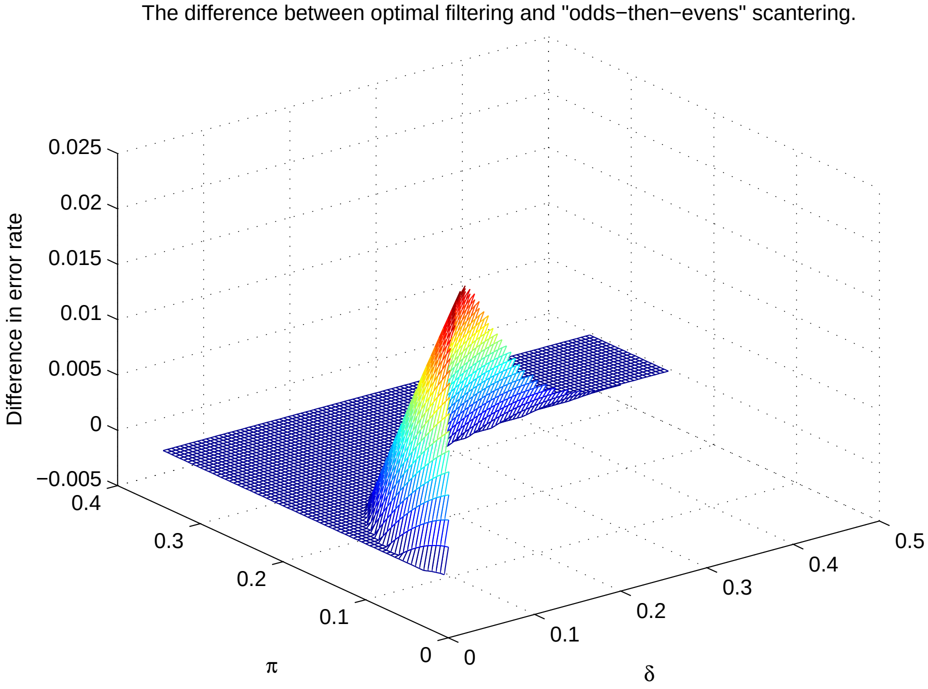

Figure 6: The actual difference between the optimal filtering error rate and the “odds-then-evens” scanning and filtering error rate. π is the transition probability of the symmetric, first order, Markov source and δ is the channel crossover probability. Only values for which $\delta < f(\pi)$ are shown.

The analogous 'prediction-filtering transformation' is

$$\hat{X}_t^F(y^t, u_t) = a_j \in A \quad \text{if} \quad \sum_{i=0}^{j-1} \sum_{s: s(y_t)=a_i} F_t(y^{t-1})[s] \leq u_t < \sum_{i=0}^{j} \sum_{s: s(y_t)=a_i} F_t(y^{t-1})[s], \tag{65}$$

where the subscript i reflects some enumeration of A. Under the above definitions, [19, Theorem 4] states that for all n, $x^n \in A^n$ and any predictor F,

$$EL_{\hat{X}^F}(x^n, Y^n) = EL'_F(Y^n), \tag{66}$$

where $L_{\hat{X}^F}(x^n, Y^n)$ is the cumulative loss of the filter under the original loss function l and $L'_F(Y^n)$ is the cumulative loss of the predictor under a *modified loss function* l', which depends on the original loss l and the channel crossover probabilities.

This result can be used for universal filtering, under invertible discrete memoryless channels, in the following way. For each finite set of filters, construct the corresponding set of predictors, then use the well known results in universal prediction in order to construct a universal predictor for that set. Finally, construct the universal filter using the "inverse" prediction-filtering transformation. Analogously, the results on universal finite set scandiction given in [12] can be used to construct universal scanner-filter pairs. Note, however, that the modified loss function l' may be much more complex to handle compared to the original one. For example, it may not be a function of the difference $x_t - F_t$, even if the original loss function is. Nevertheless, the results in [12] apply to any bounded loss function, and thus can be utilized.

4 Scandiction of Noisy Data Arrays

In this section, we consider a scenario similar to that of Section 3, only now, for each t, the data Y_{Ψ_t} is not available in the estimation of X_{Ψ_t}, namely, $F_t = F_t(Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}})$, as opposed to $\tilde{F}_t = \tilde{F}_t(Y_{\Psi_1}, \ldots, Y_{\Psi_t})$ in the filtering scenario. We refer to this scenario as "noisy scandiction", analogous to the noisy prediction problems discussed in [14] and [15].

We first assume the joint probability distribution of the underlying field and noisy observations, Q, is known, and examine the settings of Gaussian fields under squared error loss and binary fields under Hamming loss. In these cases, we characterize the noisy scandictability and the achieving scandictors in terms on the "clean" scandictability of the noisy data. We then consider universal scandiction for the noisy setting, show that this is indeed possible for finite scandictor set and for the class of all stationary binary fields corrupted by binary noise. Finally, we derive bounds on the excess loss when non optimal scanners are used (yet, with the optimal predictor for each scan).

4.1 Noisy Scandictability

Throughout this section, it will be beneficial to consider also the *clean* scandictability as defined in [11, Definition 2], that is, when the scandictor is judged with respect to the same random field it observes. Thus, for (X, Y) governed by the probability measure Q, Q_Y denotes the marginal measure of $\{Y\}$, and therefore $U(l, Q_Y)$ refers to the clean scandictability of Y, i.e.,

$$L_{(\Psi,F)}(y_B) = \sum_{t=1}^{|B|} l(y_{\Psi_t}, F_t(y_{\Psi_1}, \ldots, y_{\Psi_{t-1}})), \tag{67}$$

and

$$U(l, Q_Y) = \lim_{n\to\infty} \inf_{(\Psi,F)} E_{Q_Y} \frac{1}{|B|} L_{(\Psi,F)}(Y_B). \tag{68}$$

As mentioned earlier, in this section we relate the noisy scandictability, $\bar{U}(l, Q)$, to the clean scandictability of the noisy field, $U(l, Q_Y)$. This relation can be used to derive bounds on $\bar{U}(l, Q)$ using the bounds on $U(l, Q_Y)$ derived in [11]. However, this should be done carefully. For example, the lower and upper bounds given in [11, Theorem 9] are applicable only when X has an autoregressive representation (with respect to some scandictor) with independent innovations. Unfortunately, $Y = X + N$ does not necessarily have this representation, and the bounds do not apply to Y in a straightforward manner.[3] Yet, a simple generalization of the lower bound in [11], valid for arbitrarily distributed random fields, can be derived using the same method used in the proof of Theorem 2. To this end, we briefly describe this generalization.

Let

$$B(P) = \min_{\hat{y}} \sum_{y} l(y, \hat{y}) P(y), \tag{69}$$

and further define

$$\gamma(d) = \max\{H(P) : B(P) \leq d\}. \tag{70}$$

Similarly as in Section 3.1, denote by $\bar{\gamma}(d)$ the upper concave envelope of $\gamma(d)$.

Corollary 8. *For any random field Y_B and any scandictor (Ψ, F) for Y_B,*

$$\bar{\gamma}(L_{(\Psi,F)}(Y_B)) \geq \frac{1}{|B|} H(Y_B). \tag{71}$$

[3] Note that the restriction to autoregressive fields is merely technical, i.e., it facilitates the proof of the lower bound in the sense that a weak AEP-like theorem is required. The essence of the lower bound, however, which is a volume preservation argument, is valid for non autoregressive fields as well.

Proof. The proof is similar to that of Theorem 2. We have,

$$\begin{aligned} H(Y_B) &= H(\Psi(Y_B)) \\ &= \sum_{t=1}^{|B|} H(Y_{\Psi_t}|Y^{\Psi_{t-1}}) \\ &\leq \sum_{t=1}^{|B|} \sum_{y^{\Psi_{t-1}}} \gamma\left(E_{Q_B}\left\{l\left(Y_{\Psi_t}, F_t(y^{\Psi_{t-1}})\right)|Y^{\Psi_{t-1}} = y^{\Psi_{t-1}}\right\}\right) P(y^{\Psi_{t-1}}) \\ &\leq |B|\bar{\gamma}\left(\frac{1}{|B|}E_{Q_B}L_{(\Psi,F)}(Y_B)\right). \end{aligned} \tag{72}$$

□

The lower bound in Corollary 8 strengthens the bound in [11, Theorem 9] since it applies to general loss functions, arbitrarily distributed random fields, and is non-asymptotic. When $A = \mathbb{R}$ and the loss function is of the form $l(x, F) = \rho(x - F)$, where $\rho(z)$ is monotonically increasing for $z > 0$, monotonically decreasing for $z > 0$, satisfies $\rho(0) = 0$ and $\int e^{-s\rho(z)}dz < \infty$ for every $s > 0$, the above bound coincides with that of [11]. In that case, $\bar{\gamma}(d) = \gamma(d)$, which is in turn the *one sided Fenchel-Legendre transform* of the *log moment generating function* associated with ρ (See [11, Section III] for the details). For example, when $\rho(z) = z^2$, we have, $\gamma(d) = \frac{1}{2}\ln(2\pi e d)$, $d > 0$ and $\gamma^{-1}(h) = \frac{1}{2\pi e}e^{2h}$, $h > 0$. Similar results can be derived for binary alphabet, thus, when $\rho(z)$ is the Hamming loss function, $\gamma(d) = h_b(d)$.

We now turn to discuss the noisy scnadictability, $\bar{U}(l, Q)$. The following lemma, proved in Appendix A.2, describes the noisy scandictability for any additive white noise channel model and the squared error loss function, $l_s(\cdot)$, in terms of the clean scandictability of Y, and gives the optimal scandictor.

Lemma 9. *Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}^2}$ be a random field governed by a probability measure Q such that $Y_t = X_t + N_t$, where N_t, $t \in \mathbb{Z}^2$, are i.i.d. random variables with $Var(N_t) = \sigma_N^2 < \infty$. Then*

$$\bar{U}(l_s, Q) = U(l_s, Q_Y) - \sigma_N^2. \tag{73}$$

Furthermore, $\bar{U}(l_s, Q)$ is achieved by the scandictor which achieves $U(l_s, Q_Y)$.

Actually, Lemma 9 is only scarcely related to scanning. It merely states that in the prediction of a process based on its noisy observations, under the additive model stated above and squared error loss, the optimal predictor is one which disregards the noise, and attempts to predict the next *noisy* outcome. Similar results for binary processes through a BSC were given in [16] and will be discussed later.

Finally, we mention that the method used in the proof of Lemma 9 is specific for the square error loss function. For a general loss function, one can use conditional expectation in order to compute the noisy scandictability, under a *modified* loss function ρ. Specifically, for a random field X, denote by $\sigma(X_{V_n})$ the smallest sigma algebra with respect to which X_{V_n} is measurable. Let Ψ_n denote a scanner for V_n and denote by $\mathcal{F}_t^{\Psi_n}$ the *information available to the scandictor at the t'th step*, that is

$$\mathcal{F}_t^{\Psi_n} = \sigma\left(Y_{\Psi_1}, Y_{\Psi_2(Y_{\Psi_1})}, \ldots, Y_{\Psi_t(Y_{\Psi_1}^{\Psi_{t-1}})}\right). \tag{74}$$

Note that the set of sites $\Psi_1, \Psi_2, \ldots, \Psi_t$ is itself random, yet for each t, Ψ_t is $\mathcal{F}_{t-1}^{\Psi_n}$ measurable (if Ψ is random, namely, it uses additional independent random variables, the definition of $\mathcal{F}_t^{\Psi_n}$ is altered accordingly). Hence, the filtration $\{\mathcal{F}_t^{\Psi_n}\}_{t=1}^{|V_n|}$ represents the gathered knowledge at the scandictor. We have

$$\begin{aligned} E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \rho\left(F_t - Y_{\Psi_t}\right) &= E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} E_{Q_{B_n}} \left\{ \rho\left(F_t - X_{\Psi_t} - N_{\Psi_t}\right) \Big| \mathcal{F}_{t-1}^{\Psi}, \sigma(X_{\Psi_t}) \right\} \\ &= E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \tilde{\rho}\left(F_t - X_{\Psi_t}\right), \end{aligned} \tag{75}$$

for some $\tilde{\rho}$. Thus, if $l(X_{\Psi_t}, F_t)$ is the required loss function in the noisy prediction problem of $\{X\}$, one has to seek a function $\rho(\cdot,\cdot)$ such that $\tilde{\rho}(x_{\Psi_t}, F_t) = l(x_{\Psi_t}, F_t)$ for all x_{Ψ_t} and F_t. If such a function is found, then surely $E\rho(Y_{\Psi_t}, F_t) = El(X_{\Psi_t}, F_t)$ and the optimal scandictor for the noisy prediction problem is the one which is optimal for the clean prediction problem of $\{Y\}$ under ρ. While this is simple for the squared error loss function and additive noise (choose $\rho(y - F) = (y - F)^2 - \sigma_N^2$), or Hamming loss and BSC (choose $\rho(y, F) = \frac{l_H(y,F) - \delta}{1 - 2\delta}$) this is not always the case for a general loss function. It is also important to note that in the case of white noise considred in this paper, the condition on the modified loss function ρ can be stated in a single letter expresion, namly, if $l(X, F)$ is the required loss function for the noisy scandiction problem, ρ should satisfy $E\{\rho(Y, F)|\sigma(X)\} = l(X, F)$.

4.1.1 Gaussian Random Fields

Let both X and N be Gaussian random fields, where the components of N are i.i.d. and independent of X. That is, Y is the output of an AWGN channel, with a Gaussian input X. In this scenario, similarly to the clean one, the noisy scandictability is known exactly and is given by a single letter expression.

Before we proceed, several definitions are required. For any $t \in \mathbb{Z}^2$ and $V \subseteq \mathbb{Z}^2$, denote by $\hat{X}_t(V)$ the best linear predictor of X_t given $\{X_{t'}\}_{t' \in V}$. A subset $S \subseteq \mathbb{Z}^2$ is called a *half* *plane* if it is closed to addition and satisfies $S \cup (-S) = \mathbb{Z}^2$ and $S \cap (-S) = \{0\}$. For example, $S_{\text{lex}} = \{(m,n) \in \mathbb{Z}^2 : [m > 0] \text{ or } [m = 0, n \geq 0]\}$ is a half plane. Let X be a wide sense stationary random field and denote by g the density function associated with the absolutely continuous component in the Lebesgue decomposition of its spectral measure. Then, for any half plane S, we have [26, Theorem 1],

$$\begin{aligned} E\left(X_0 - \hat{X}_0(-S \setminus \{0\})\right)^2 &= \exp\left\{\frac{1}{4\pi^2}\int_{[0,2\pi)^2} \ln g(\lambda) d\lambda\right\} \\ &\overset{\triangle}{=} \sigma_u^2(X). \end{aligned} \tag{76}$$

We can now state the following corollary, regarding the noisy scandictability in the Gaussian regime and squared error loss, which is a direct application of Lemma 9 and the results of [11, Section IV].

Corollary 10. *Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}^2}$ be a random field governed by a probability measure Q such that $Y_t = X_t + N_t$, where X is a stationary Gaussian random field, N_t, $t \in \mathbb{Z}^2$, is an AWGN, independent of $\{X_t\}_{t\in\mathbb{Z}^2}$. Then, the noisy scandictability of Q under the squared error loss is given by*

$$\bar{U}(l_s, Q) = \sigma_u^2(Y) - \sigma_N^2. \tag{77}$$

Furthermore, $\bar{U}(l_s, Q)$ is asymptotically achieved by a scandictor which scans (X_t, Y_t) according to the total order defined by any half-plane S and applies the corresponding best linear predictor for the next outcome of Y.

For any stationary Gaussian *process* X, it has been shown by Kolmogorov (see for example [27]) that the entropy rate is given by

$$H_*^X = \frac{1}{2}\ln(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln g(\lambda) d\lambda. \tag{78}$$

Thus, using the one-dimensional analogue of (76), for a stationary Gaussian process X we have,

$$H_*^X = \frac{1}{2}\ln(2\pi e \sigma_u^2(X)). \tag{79}$$

In fact, (79) applies for stationary Gaussian random fields as well. Thus, we have,

$$\begin{aligned} \bar{U}(l_s, Q) &= \sigma_u^2(Y) - \sigma_N^2 \\ &= \frac{1}{2\pi e}e^{2H_*^Y} - \frac{1}{2\pi e}e^{2H^N}, \end{aligned} \tag{80}$$

where H_*^Y is the entropy rate of Y and H^N is the entropy of each N_t. From the entropy power inequality [23, pp. 496], we have,

$$\frac{1}{2\pi e}e^{2H_*^Y} \geq \frac{1}{2\pi e}e^{2H^N} + \frac{1}{2\pi e}e^{2H_*^X}, \tag{81}$$

thus, as expected, the noisy scandictability given in Corollary 10 (and (80)) is at least as small the clean scandictability of X, that is, with no noise at all. In most of the interesting cases, however, (81) is a strict inequality. In fact, as mentioned in [28], (81) is achieved with equality only when both X and N are Gaussian and have *proportional spectra*. Consequently, unless X is white, Corollary 10 is non-trivial.

4.1.2 Binary Random Fields

In this case, the results of [14] and [16] shed light on the optimal scandictor. Therein, it was shown that for a binary prediction problem, i.e., where $\{X_t\}$ is a binary source passed through a BSC with cross over probability $\delta < \frac{1}{2}$, and $\{Y_t\}$ is the channel output, the more likely outcome for the clean bit is also the more likely outcome for the noisy bit. Thus, the optimal predictor in the Hamming sense for the next clean bit (based on the noisy observations) might as well use the same strategy as if it tries to predict the next noisy bit. Consequently, the optimal *scandictor* in the noisy setting is the one which is optimal for $\{Y\}$, and the results of [11, Section V] apply.

The following proposition relates the scandictability of a binary noise-corrupted process $\{Y\}$, judged with respect to the clean binary process $\{X\}$, to its clean scandictability.

Proposition 11. *Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}^2}$ be a binary random field governed by a probability measure Q such that $\{Y_t\}$ is the output of a binary memoryless symmetric channel with cross over probability δ and input $\{X_t\}$. Then,*

$$\bar{U}(l_H, Q) = \frac{U(l_H, Q_Y) - \delta}{1 - 2\delta}, \tag{82}$$

where l_H is the Hamming loss function. Furthermore, $\bar{U}(l_H, Q)$ is achieved by the scandictor which achieves $U(l_H, Q_Y)$.

Note that indeed $U(l_H, Q_Y) \geq \delta$ as Y is the output of a BSC with crossover probability δ.

Proof (Proposition 11). Let $\{B_n\}_{n\geq 1}$ be any sequence of elements in $\mathcal{V}$, satisfying $R(B_n) \to \infty$. We have,

$$\begin{aligned}\bar{U}(l_H, Q_{B_n}) &= \inf_{(\Psi,F)\in\mathcal{S}(B_n)} E_{Q_{B_n}} \frac{1}{|B_n|}\sum_{t=1}^{|B_n|} l_H(X_{\Psi_t}, F_t(Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}})) \\ &= \inf_{(\Psi,F)\in\mathcal{S}(B_n)} \frac{1}{|B_n|}\sum_{t=1}^{|B_n|} P\left(F_t(Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}}) \neq X_{\Psi_t}\right), \qquad (83)\end{aligned}$$

and, analogously,

$$U(l_H, Q_{Y,B_n}) = \inf_{(\Psi,F)\in\mathcal{S}(B_n)} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} P\left(F_t(Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}}) \neq Y_{\Psi_t}\right). \tag{84}$$

Denoting by Z_t the channel noise at time t, and abbreviating $F_t(Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}})$ by F_t, we have

$$\begin{aligned} P\left(F_t \neq Y_{\Psi_t}\right) &= P\left(F_t \neq Y_{\Psi_t}, Z_{\Psi_t} = 1\right) + P\left(F_t \neq Y_{\Psi_t}, Z_{\Psi_t} = 0\right) \\ &= P\left(F_t = X_{\Psi_t}, Z_{\Psi_t} = 1\right) + P\left(F_t \neq X_{\Psi_t}, Z_{\Psi_t} = 0\right) \\ &= \left(1 - P\left(F_t \neq X_{\Psi_t}\right)\right)\delta + P\left(F_t \neq X_{\Psi_t}\right)(1-\delta). \end{aligned} \tag{85}$$

Namely, for $\delta < \frac{1}{2}$, the optimal strategy for predicting Y_{Ψ_t} based on $Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}}$ and the optimal strategy for predicting X_{Ψ_t} based on $Y_{\Psi_1}, \ldots, Y_{\Psi_{t-1}}$ are identical, and, in addition,

$$P\left(F_t \neq X_{\Psi_t}\right) = \frac{P\left(F_t \neq Y_{\Psi_t}\right) - \delta}{1 - 2\delta}. \tag{86}$$

Substituting (86) into (83) and taking $n \to \infty$ completes the proof. □

4.2 Universal Scandiction in the Noisy Scenario

Section 4.1 dealt with the actual value of the best achievable performance in the noisy scandiction scenario. However, it is also interesting to investigate the universal setting in which one seeks a predictor which does not depend on the joint probability measure of $\{(X, Y)\}$, yet performs asymptotically as well as a one matched to this measure. The problem of universal scandiction in the noiseless scenario was dealt with in [12]. Herein, we show that it is possible to construct universal scandictors in the noisy setting as well (similar to universal scanning and filtering in Section 3.3). First, we show that it is possible to compete successfully with any finite set of scandictors, and present a universal scandictor for this setting. We then show that with a proper choice of a set of scandictors, it is possible to (universally) achieve $\bar{U}(l, Q)$, i.e., the noisy scandictability, for any spatially stationary random filed (X, Y).

At the basis of the results of [12] stands the exponential weighting algorithm, originally derived by Vovk in [29]. In [29], Vovk considered a general set of *experts* and introduced the exponential weighting algorithm in order to compete with the best expert in the set. In this algorithm, each expert is assigned with a weight, according to its past performance. By decreasing the weight of poorly performing experts, hence preferring the ones proved to perform well thus far, one is able to compete with the best expert, having neither any *a priori* knowledge on the input sequence nor which expert will perform the best. It is clear that the essence of this algorithm is the use of the cumulative losses incurred by each expert to construct a probability measure on the experts, which is later used to choose an expert for the next action. However, when the clean data X is not known to the sequential algorithm, it is impossible to calculate the cumulative losses of the experts precisely. Nevertheless, as Weissman and Merhav show in [15], using an unbiased estimate $\hat{X}_t(Y_t)$ of X_t results in sufficiently accurate estimates of the cumulative losses of the experts, which in turn can be used by the exponential weighting algorithm. Hence, the framework derived in [12] can then be used to suggest universal scandictors for the noisy setting as well.

Consider a random field (X_B, Y_B) where X is binary and Y is either binary (e.g., the output of a BSC whose input is Y) or real valued (e.g., X through a Gaussian noise channel). For a loss function $l : \{0,1\} \times [0,1] \to [0,\infty]$ we define, similarly to [15],

$$l_0(\cdot) \overset{\triangle}{=} l(0,\cdot) \quad \text{and} \quad l_1(\cdot) \overset{\triangle}{=} l(1,\cdot). \tag{87}$$

Assume (Ψ, F) is a scandictor for B. Then, for any $t \le |B|$, we have

$$\begin{aligned} L_{(\Psi,F)}(x_B, y_B)_t &= \sum_{i=1}^{t} l(F_i(y_{\Psi_1}^{\Psi_{i-1}}), x_{\Psi_i}) \\ &= \sum_{i=1}^{t} \left[(1 - x_{\Psi_i}) l_0(F_i(y_{\Psi_1}^{\Psi_{i-1}})) + x_{\Psi_i} l_1(F_i(y_{\Psi_1}^{\Psi_{i-1}})) \right]. \end{aligned} \tag{88}$$

Clearly, $L_{(\Psi,F)}(x_B, y_B)_t$ depends on x_B and is not known to the sequential algorithm. Let $h(y_{\Psi_i})$ be an unbiased estimate for x_{Ψ_i}. For example, when Y is the output of a BSC with input X we may choose

$$h(y_{\Psi_i}) = \frac{y_{\Psi_i} - \delta}{1 - 2\delta}. \tag{89}$$

Define

$$\hat{L}_{(\Psi,F)}(y_B)_t = \sum_{i=1}^{t} \left[(1 - h(y_{\Psi_i})) l_0(F_i(y_{\Psi_1}^{\Psi_{i-1}})) + h(y_{\Psi_i}) l_1(F_i(y_{\Psi_1}^{\Psi_{i-1}})) \right], \tag{90}$$

and

$$\begin{aligned} \Delta_{(\Psi,F)}(x_B, y_B)_t &\overset{\triangle}{=} L_{(\Psi,F)}(x_B, y_B)_t - \hat{L}_{(\Psi,F)}(y_B)_t \\ &= \sum_{i=1}^{t} \left(h(y_{\Psi_i}) - x_{\Psi_i}\right) l_0(F_i(y_{\Psi_1}^{\Psi_{i-1}})) + \sum_{i=1}^{t} \left(x_{\Psi_i} - h(y_{\Psi_i})\right) l_1(F_i(y_{\Psi_1}^{\Psi_{i-1}})). \end{aligned} \tag{91}$$

Similarly to [15], we assume that the noise field N_B is of independent components and that for each $i \in B$, $Y_i \in \sigma(N_i)$, i.e., the noise component at site i affects the observation at that site alone. In Appendix A.3, we show that for any image x_B and any scandictor (Ψ, F) for

B, $\left(\Delta_{(\Psi,F)}(x_B, y_B)_t, \mathcal{F}_t^\Psi\right)$ is a zero mean martingale. As a result, for any scandictor (Ψ, F), image x_B and t we have

$$EL_{(\Psi,F)}(x_B, Y_B)_t = E\hat{L}_{(\Psi,F)}(Y_B)_t, \tag{92}$$

namely, $\hat{L}_{(\Psi,F)}(Y_B)_t$ is an unbiased estimator for $L_{(\Psi,F)}(x_B, Y_B)_t$. The universal algorithm for scanning and prediction in the noisy scenario will thus use $\hat{L}_{(\Psi,F)}(Y_B)_t$ instead of $L_{(\Psi,F)}(x_B, Y_B)_t$, which is unknown. More specifically, similarly to the algorithm proposed in [12], the algorithm divides the data array to be scandicted to blocks of size $m(n) \times m(n)$, then scans the data in a (fixed) block-wise order, where each block is scandicted using a scandictor chosen at random from the scandictors set, according to the distribution $\hat{P}_i\left(j|\{\hat{L}_{j,i}\}_{j=1}^{\lambda}\right)$,

$$\hat{P}_i\left(j|\{\hat{L}_{j,i}\}_{j=1}^{\lambda}\right) = \frac{e^{-\eta \hat{L}_{j,i}}}{\sum_{j=1}^{\lambda} e^{-\eta \hat{L}_{j,i}}}, \tag{93}$$

where $\hat{L}_{j,i} = \sum_{m=0}^{i-1} \hat{L}_{(\Psi,F)_j}(y^m)$, the estimated cumulative loss of the scandictor $(\Psi, F)_j$ after scandicting i blocks of data, when $(\Psi, F)_j$ is restarted after each block, and λ is the cardinality of the set of scandictors, $\mathcal{F}_m$.[4] Note the subscript m in $\mathcal{F}_m$, as in order to scandict a data array of size $n \times n$, the universal algorithm discussed herein uses the scandictors with which it competes, but only on blocks of size $m \times m$.

The following proposition gives an upper bound on the redundancy of the algorithm when competing with a finite set of scandictors, each operating block-wise on the data array.

Proposition 12. *Let $EL_{alg}(x_{V_n}, Y_{V_n})$ be the expected (with respect to the noisy random field as well as the randomization in the algorithm) cumulative loss of the proposed algorithm on Y_{V_n}, when the underlying clean array is x_{V_n} and the noisy field is of independent components with $Y_i \in \sigma(N_i)$ for each $i \in V_n \subset \mathbb{Z}^2$. Let $EL_{min}(x_{V_n}, Y_{V_n})$ denote the expected cumulative loss of the best scandictor in $\mathcal{F}_m$, operating block-wise on Y_{V_n}. Assume $|\mathcal{F}_m| = \lambda$, then*

$$EL_{alg}(x_{V_n}, Y_{V_n}) - EL_{min}(x_{V_n}, Y_{V_n}) \leq m(n)(n + m(n))\sqrt{\ln \lambda}\frac{l_{max}}{\sqrt{2}}. \tag{94}$$

[4] To be consistent with the notation of [12], the same notation is used for both a filtration and a scandictor set. The difference should be clear from the context.

Proof. By (92) and [12, Proposition 3], for any x_{V_n} we have

$$\begin{aligned} EL_{alg}(x_{V_n}, Y_{V_n}) - \min_{(\Psi,F)\in\mathcal{F}_m} EL_{(\Psi,F)}(x_{V_n}, Y_{V_n}) \\ &= E\hat{L}_{alg}(Y_{V_n}) - \min_{(\Psi,F)\in\mathcal{F}_m} E\hat{L}_{(\Psi,F)}(Y_{V_n}) \\ &\leq E\hat{L}_{alg}(Y_{V_n}) - E\min_{(\Psi,F)\in\mathcal{F}_m} \hat{L}_{(\Psi,F)}(Y_{V_n}) \\ &= E\left\{\hat{L}_{alg}(Y_{V_n}) - \min_{(\Psi,F)\in\mathcal{F}_m} \hat{L}_{(\Psi,F)}(Y_{V_n})\right\} \\ &\leq m(n)(n+m(n))\sqrt{\ln\lambda}\frac{l_{max}}{\sqrt{2}}. \end{aligned} \tag{95}$$

□

Proposition 12 is the basis for the main result in this sub-section, a universal scandictor which competes successfully with any finite set of scandictors for the noisy scenario.

Theorem 13. *Let (X,Y) be a stationary random field with a probability measure Q. Assume that for each $i \in \mathbb{Z}^2$, Y_i is the output of a memoryless channel whose input is X_i. Let $\mathcal{F} = \{\mathcal{F}_n\}$ be an arbitrary sequence of scandictor sets, where $\mathcal{F}_n$ is a set of scandictors for V_n and $|\mathcal{F}_n| = \lambda < \infty$ for all n. Then, there exists a sequence of scandictors $\{(\hat{\Psi},\hat{F})_n\}$, independent of Q, for which*

$$\liminf_{n\to\infty} E_{Q_{V_n}} E\frac{1}{|V_n|}L_{(\hat{\Psi},\hat{F})_n}(X_{V_n}, Y_{V_n}) \leq \liminf_{n\to\infty} \min_{(\Psi,F)\in\mathcal{F}_n} E_{Q_{V_n}}\frac{1}{|V_n|}L_{(\Psi,F)}(X_{V_n}, Y_{V_n}) \tag{96}$$

for any $Q \in \mathcal{M}_S(\Omega)$, where the inner expectation in the l.h.s. of (96) *is due to the possible randomization in $(\hat{\Psi},\hat{F})_n$.*

The proof of Theorem 13 follows the proof of [12, Theorem 2] verbatim.

It is now possible to show the existence of a universal scandictor for any stationary random field in the noisy scandiction setting. Herein, we include only the setting where X is binary and Y is the output of a BSC. In this case, the scandictor is twofold-universal, namely, it does not depend on the channel crossover probability either. Extending the results to real-valued noise is possible using the methods introduced in [16] (although the universal predictor does depend on the channel characteristics) and will be discussed later.

Theorem 14. *Let X be a stationary random field over a finite alphabet A and a probability measure Q. Let Y be the output of a BSC whose input is X and whose crossover probability δ. Let the prediction space D be either finite or bounded (with $l(x,F)$ then being Lipschitz in* *its second argument). Then, there exists a sequence of scandictors* $\{(\Psi, F)_n\}$, *independent of* Q *and of* δ, *for which*

$$\lim_{n\to\infty} E_{Q_{V_n}} E \frac{1}{|V_n|} L_{(\Psi,F)_n}(X_{V_n}, Y_{V_n}) = \tilde{U}(l, Q) \tag{97}$$

for any $Q \in \mathcal{M}_S(\Omega)$, *where the inner expectation in the l.h.s. of* (97) *is due to the possible randomization in* $(\Psi, F)_n$.

Similar to [16, Section A] and the proof of [12, Theorem 6], in the case of binary input and binary-valued noise it is possible to take the set of scandictors with which we compete as the set of *all possible scandictors for an* $m(n) \times m(n)$ block. The proof thus follows directly from the proof of [12, Theorem 6].

As for continuous-valued observations, it is quite clear that the set of all possible scandictors for an $m(n) \times m(n)$ block is far too rich to compete with (note that this is since the number of *predictors* is too large). A complete discussion is available in [16, Section B]. However, Weissman and Merhav do offer a method for successfully achieving the Bayes envelope for this setting, by introducing a much smaller set of predictors, which on one hand *includes* the best kth order Markov predictor, yet on the other hand is not too rich, in the sense that the redundancy of the exponential weighting algorithm tends to zero when competing *with an* ϵ*-grid of it.* Since presenting a universal *scandictor* for this scenario will mainly include a repetition of the many details discussed in [16], we do not include it here.

4.3 Bounds on the Excess Loss for Non-Optimal Scandictors

Analogously to the scanning and filtering setting discussed in Section 3, and the clean prediction setting discussed in [12], it is interesting to investigate the excess loss incurred when non optimal scandictors are used in the noisy scandiction setting. Unlike the scanning and filtering setting, where the excess loss bounds were not a straightforward extensions of the results in [12], in the noisy scandiction scenario this problem can be quite easily tackled using the results of [12] and modified loss functions.

We briefly state the results of [12] in this context. The scenario considered therein is that of predicting the next outcome of a binary source, with $D = [0, 1]$ as the prediction space. ϕ_ρ denotes the Bayes envelope associated with the loss function ρ, i.e.,

$$\phi_\rho(p) = \min_{q\in[0,1]} [(1-p)\rho(0, q) + p\rho(1, q)]. \tag{98}$$

Similarly to (54), define

$$\epsilon_\rho = \min_{\alpha,\beta} \max_{0\leq p\leq 1} |\alpha h_b(p) + \beta - \phi_\rho(p)|. \tag{99}$$

Note that although the definitions of $\phi_\rho(p)$ and ϵ_ρ refer to the binary scenario, the result below holds for larger alphabets, with ϵ_ρ defined as in (99), with the maximum ranging over the simplex of all distributions on the alphabet, and $h(p)$ (replacing $h_b(p)$) and $\phi_\rho(p)$ denoting the entropy and Bayes envelope of the distribution p, respectively. In [12], it is shown that if X_B is an arbitrarily distributed binary random field, then, for any scan Ψ,

$$\left|\alpha_\rho \frac{1}{|B|} H(X_B) + \beta_\rho - E_{Q_B} \frac{1}{|B|} L_{(\Psi,F^{opt})}(X_B)\right| \leq \epsilon_\rho, \tag{100}$$

where α_ρ and β_ρ are the achievers of the minimum in (99).

As mentioned earlier, if $\rho(Y,F)$ is some loss function for the "clean" prediction problem of $\{Y\}$, the noisy process, then,

$$E\left\{\rho(Y,F)|\sigma(X)\right\} = \tilde{\rho}(X,F) \tag{101}$$

for some $\tilde{\rho}$. Assuming a suitable ρ is found (i.e., $\tilde{\rho} = l$), we have, for any scan Ψ,

$$\begin{aligned}\left|\alpha_\rho \frac{1}{|B|} H(Y_B) + \beta_\rho - \frac{1}{|B|} E_{Q_B} L^l_{\Psi,F^{opt}}(X_B,Y_B)\right| \\ &= \left|\alpha_\rho \frac{1}{|B|} H(Y_B) + \beta_\rho - \frac{1}{|B|} E_{Q_B} L^\rho_{\Psi,F^{opt}}(Y_B)\right| \\ &\leq \epsilon_\rho, \end{aligned} \tag{102}$$

where $\frac{1}{|B|}E_{Q_B}L^l_{\Psi,F^{opt}}(X_B,Y_B)$ is the normalized expected cumulative loss in optimally predicting X_{Ψ_t} based on $Y_1^{\Psi_{t-1}}$, under the loss function l, $\frac{1}{|B|}E_{Q_B}L^\rho_{\Psi,F^{opt}}(Y_B)$ is the normalized expected cumulative loss in optimally predicting Y_{Ψ_t} based on $Y_1^{\Psi_{t-1}}$, under the loss function ρ, and α_ρ and β_ρ are the minimizers of ϵ_ρ as defined in (99). Hence, the following corollary applies.

Corollary 15. *Let X_B be an arbitrarily distributed binary field. Assume a white noise, and denote the noisy version of X_B by Y_B. Let $D = [0,1]$ be the prediction space and $l : \{0,1\}\times D \to \mathbb{R}$ be any loss function. Then, for any scan Ψ,*

$$\left|\frac{1}{|B|} E_{Q_B} L^l_{\Psi,F^{opt}}(X_B,Y_B) - \bar{U}(l,Q_B)\right| \leq 2\epsilon_\rho, \tag{103}$$

when ρ is a loss function such that $E\left\{\rho(Y,F)|\sigma(X)\right\} = l(X,F)$ for any F.

Example 3 (*BSC and Hamming Loss*). In the case of binary input, BSC with crossover probability δ and Hamming loss $l_H(\cdot,\cdot)$, it is not hard to show that

$$\rho(y,F) = \frac{l_H(y,F)-\delta}{1-2\delta}. \tag{104}$$

Hence,

$$\phi_\rho(p) = \frac{\phi_{l_H}(p)-\delta}{1-2\delta} \tag{105}$$

and

$$\epsilon_\rho = \frac{1}{1-2\delta}\epsilon_{l_H}, \tag{106}$$

where $\epsilon_{l_H} = 0.08$ as mentioned in [12]. The above bound on the excess loss can also be computed directly, without using Corollary 15, as for any scan Ψ, the normalized cumulative prediction errors are given by

$$\frac{1}{|B|}E_{Q_B}L^{l_H}_{\Psi,F}(X_B,Y_B) = \frac{1}{|B|}\sum_{t=1}^{|B|}P\left(F_t(Y_{\Psi_1},\dots,Y_{\Psi_{t-1}}) \neq X_{\Psi_t}\right) \tag{107}$$

for the noisy scenario, and

$$\frac{1}{|B|}E_{Q_B}L^{l_H}_{\Psi,F}(Y_B) = \frac{1}{|B|}\sum_{t=1}^{|B|}P\left(F_t(Y_{\Psi_1},\dots,Y_{\Psi_{t-1}}) \neq Y_{\Psi_t}\right) \tag{108}$$

for the (clean) prediction of Y_B. Hence, using (86), for any scan Ψ we have,

$$\begin{aligned}
&\left|\frac{1}{|B|}E_{Q_B}L^{l_H}_{\Psi,F^{opt}}(X_B,Y_B) - \bar{U}(l_H,Q_B)\right| \\
&\quad= \left|\frac{\frac{1}{|B|}E_{Q_B}L^{l_H}_{\Psi,F^{opt}}(Y_B)-\delta}{1-2\delta} - \frac{U(l_H,Q_{Y,B})-\delta}{1-2\delta}\right| \\
&\quad= \frac{1}{1-2\delta}\left|\frac{1}{|B|}E_{Q_B}L^{l_H}_{\Psi,F^{opt}}(Y_B) - U(l_H,Q_{Y,B})\right| \\
&\quad\leq \frac{2\epsilon_{l_H}}{1-2\delta}.
\end{aligned} \tag{109}$$

Example 4 (*Additive Noise and Squared Error*). Let Y_B be the output of an additive channel, with σ_v^2 denoting the noise variance. Let l_s be the squared error loss function. In this case,

$$E\Big\{\underbrace{(Y_{\Psi_t}-F_t(y^{\Psi_{t-1}}))^2-\sigma_v^2}_{\rho(Y_{\Psi_t},F_t(y^{\Psi_{t-1}}))}|\sigma(X_{\Psi_t})\Big\} = \underbrace{(X_{\Psi_t}-F_t(y^{\Psi_{t-1}}))^2}_{l_s(X_{\Psi_t},F_t(y^{\Psi_{t-1}}))}. \tag{110}$$

Thus, Corollary 15 applies with $\rho(Y,\hat{Y}) = (Y-\hat{Y})^2-\sigma_v^2$, and clearly $\epsilon_\rho = \epsilon_{l_s}$. Note that although Corollary 15 is stated for binary alphabet, it is not hard to generalize its result to larger alphabets, as mentioned in [12, Section 4].

4.3.1 Excess Loss Bounds Via the Continuous Time Mutual Information

The bound on the excess noisy scandiction loss given in Corollary 15 was derived using the results of [12] and modified loss functions. However, new bounds can also be derived using the same method which was used in the proof of Theorem 3, namely, the scan invariance of the mutual information and the relation to the continuous time problem. We briefly discuss how such a bound can be derived for noisy scandiction of Gaussian fields corrupted by Gaussian noise.

Using the notation of Section 3.2.1, we have

$$\begin{aligned}\mathrm{Var}(X) - \int_0^1 \mathrm{Var}(X_t|Y^t)\mathrm{d}t &= \sigma_X^2 - \sigma_N^2 \ln\left(1+\frac{\sigma_X^2}{\sigma_N^2}\right) \\ &= \sigma_N^2 g\left(\frac{\sigma_X^2}{\sigma_N^2}\right), \qquad (111)\end{aligned}$$

where

$$g(x) = x - \ln(1+x). \qquad (112)$$

Since $\sigma_X^2 \geq \mathrm{Var}\left(X_{\Psi_i}|Y_{\Psi_1}^{\Psi_{i-1}}\right)$ and $g(x)$ is monotonically increasing for $x > 0$, derivations similar to (27) lead to

$$\frac{1}{n^2}EL_{(\Psi,F^{opt})}\left(X_{V_n},Y_{V_n}\right) \leq \sigma_N^2 g\left(\frac{\sigma_X^2}{\sigma_N^2}\right) + \frac{1}{n^2}2\sigma_N^2 I\left(X_{V_n},Y_{V_n}\right). \qquad (113)$$

On the other hand, since $g(x) \geq 0$ for $x \geq 0$, we have

$$\frac{1}{n^2}EL_{(\Psi,F^{opt})}\left(X_{V_n},Y_{V_n}\right) \geq \frac{1}{n^2}2\sigma_N^2 I\left(X_{V_n},Y_{V_n}\right), \qquad (114)$$

which now can be viewed as the scanning and prediction analogue of [18, eq. (156b)]. We thus have the following corollary.

Corollary 16. *Let X_{V_n} be a Gaussian random field with a constant marginal distribution satisfying Var$(X_i) = \sigma_X^2 < \infty$ for all $i \in V_n$. Let $Y_i = X_i + N_i$, where N_{V_n} is a white Gaussian noise of variance σ_N^2, independent of X_{V_n}. Then, for any two scans Ψ^1 and Ψ^2 and their optimal predictors, we have*

$$\frac{1}{n^2}\left|EL_{(\Psi^1,F^{opt})}\left(X_{V_n},Y_{V_n}\right) - EL_{(\Psi^2,F^{opt})}\left(X_{V_n},Y_{V_n}\right)\right| \leq \sigma_N^2 g\left(\frac{\sigma_X^2}{\sigma_N^2}\right). \qquad (115)$$

Similarly as in Theorem 3, the bound in Corollary 16 has the form $\sigma_X^2\frac{g(\mathrm{SNR})}{\mathrm{SNR}}$, namely, it scales with the variance of the input. As expected, at the limit of low SNR, $\frac{g(\mathrm{SNR})}{\mathrm{SNR}} \rightarrow 0$, since regardless of the scan, one is clueless about the underlying clean symbol. In fact, it is not surprising that this behavior is common to both the filtering and the prediction scenarios. In the former, the bound value is given by (23), while in the latter it is given by (111). In both cases, the bound value is simply the difference between a continuous time filtering problem, and a discrete time filtering (or prediction, in (111)) problem. It is not hard to see that this difference tends to 0 as $\mathrm{SNR} \to 0^+$. At the limit of high SNR, $\frac{g(\mathrm{SNR})}{\mathrm{SNR}} \to 1$. Indeed, this limit corresponds to the noiseless scandiction scenario, where scanning is consequential [11].

5 Conclusion

We investigated problems in sequential filtering and prediction of noisy multidimensional data arrays. A bound on the best achievable scanning and filtering performance was derived, and the excess loss incurred when non-optimal scanners are used was quantified. In the prediction setting, a relation of the best achievable performance to that of the clean scandictability was given. In both the filtering and prediction scenarios, a special emphasis was given to the cases of AWGN and squared error loss, and BSC and Hamming loss.

Due to their sequential nature, the problems discussed in this paper are strongly related to the filtering and prediction problems where reordering of the data is not allowed (or where there is only one natural order to scan the data), such as robust filtering and universal prediction discussed in the current literature. However, the numerous scanning possibilities in the multidimensional setting add a multitude of new challenges. In fact, many interesting problems remain open. It is clear that identifying the optimal scanning methods in the widely used input and channel models discussed herein is required, as the implementation of universal algorithms might be too complex in realistic situations. Moreover, tighter upper bounds on the excess loss can be derived in order to better understand the trade-offs between non-trivial scanning methods and the overall performance. Finally, by [11], the trivial scan is optimal for scandiction of noise-free Gaussian random fields. By Corollary 10 herein, this is also the case in scandiction of Gaussian fields corrupted by Gaussian noise. Whether the same hold for scanning and *filtering* of Gaussian random fields corrupted by Gaussian noise remains unanswered.

A Appendixes

A.1 Proof of Theorem 4

The proof resembles the proof of Theorem 4. However, the derivations leading to the analogue of (27) are slightly different. For any input field X_{V_n}, we have

$$\begin{aligned}
&\frac{1}{n^2} E_{Q_{V_n}} L_{(\Psi,\tilde{F}^{opt})}\left(X_{V_n}, Y_{V_n}\right) \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n^2} \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_i}\right) \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n^2} \Bigg\{ \int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_t^{(c)}\}_{t\in[i-1,i-1+t]}\right) \mathrm{d}t \\
&\qquad\qquad - \left[\int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_t^{(c)}\}_{t\in[i-1,i-1+t]}\right) \mathrm{d}t - \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_i}\right) \right] \Bigg\} \\
&\qquad \overset{(a)}{\geq} \frac{1}{n^2}\sum_{i=1}^{n^2} \int_0^1 \mathrm{Var}\left(X_{\Psi_i} | Y_{\Psi_1}^{\Psi_{i-1}}, \{Y_t^{(c)}\}_{t\in[i-1,i-1+t]}\right) \mathrm{d}t - f^*\left(X_{V_n}, \sigma_N^2\right) \\
&\qquad = \frac{1}{n^2} 2\sigma_N^2 I\left(X_{V_n}; Y_{V_n}\right) - f^*\left(X_{V_n}, \sigma_N^2\right), \qquad (116)
\end{aligned}$$

where (a) results from the definition of $f^*()$. The rest of the proof follows similar to the proof of Theorem 4, since for any X_{V_n} and σ_N^2 it is clear that $f^*(X_{V_n}, \sigma_N^2)$ is non negative.

A.2 Proof of Lemma 9

Without loss of generality, we assume $EN_t = 0$. From Proposition 1, we have,

$$\bar{U}(\rho_s, Q) = \lim_{n\to\infty} \inf_{(\Psi,F)\in\mathcal{S}(B_n)} E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \left(X_{\Psi_t} - F_t(y_{\Psi_1}, \ldots, y_{\Psi_{t-1}})\right)^2. \qquad (117)$$

However, since $Y_{\Psi_t} = X_{\Psi_t} + N_{\Psi_t}$, and N_{Ψ_t} is independent of $N_{\Psi_{t'}}$, $t' \neq t$ and of all $\{X_t\}$, we have

$$\begin{aligned}
&E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \left(X_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) \right)^2 \\
&\qquad = E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \left(Y_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) - N_{\Psi_t} \right)^2 \\
&\qquad = E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \Bigg\{ \left(Y_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) \right)^2 \\
&\qquad\qquad -2N_{\Psi_t} \left(X_{\Psi_t} + N_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) \right) + N_{\Psi_t}^2 \Bigg\} \\
&\qquad = E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \left(Y_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) \right)^2 - \sigma_N^2. \qquad (118)
\end{aligned}$$

That is,

$$U(\rho_s, Q) = \lim_{n \to \infty} \inf_{(\Psi,F) \in \mathcal{S}(B_n)} E_{Q_{B_n}} \frac{1}{|B_n|} \sum_{t=1}^{|B_n|} \left(Y_{\Psi_t} - F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}}) \right)^2 - \sigma_N^2, \qquad (119)$$

which completes the proof.

A.3 The Martingale Property of $\left(\Delta_{(\Psi,F)}(x_B, y_B)_t, \mathcal{F}_t^\Psi\right)$

The proof follows that of [15, Lemma 1]. However, notice that due to the data-dependent scanning $\mathcal{F}_t^\Psi$ is not generated by a *fixed* set of random variables, that is, over a fixed set of sites, but by a set of t random variables which may be different for each instantiation of the random field (as for each t, Ψ_t depends on $Y_{\Psi_1}^{\Psi_{t-1}}$). Yet, the expectation will always be with respect to the random variables *seen so far*.

By (91),

$$\Delta_{(\Psi,F)}(x_B, y_B)_t = \sum_{i=1}^{t} \left(h(y_{\Psi_i}) - x_{\Psi_i} \right) l_0(F_i(y_{\Psi_1}^{\Psi_{i-1}})) + \sum_{i=1}^{t} \left(x_{\Psi_i} - h(y_{\Psi_i}) \right) l_1(F_i(y_{\Psi_1}^{\Psi_{i-1}})). \quad (120)$$

Defining

$$m_t \triangleq \sum_{i=1}^{t} \left(h(Y_{\Psi_i}) - x_{\Psi_i} \right) l_0(F_i(Y_{\Psi_1}^{\Psi_{i-1}})), \qquad (121)$$

we have,

$$\begin{aligned}
E\left\{m_{t+1}|\mathcal{F}_t^{\Psi}\right\} &= E\left\{\sum_{i=1}^{t+1}\left(h(Y_{\Psi_i})-x_{\Psi_i}\right)l_0(F_i(Y_{\Psi_1}^{\Psi_{i-1}}))|\mathcal{F}_t^{\Psi}\right\} \\
&= E\left\{\left(h(Y_{\Psi_{t+1}})-x_{\Psi_{t+1}}\right)l_0(F_{t+1}(Y_{\Psi_1}^{\Psi_t}))|\mathcal{F}_t^{\Psi}\right\} \\
&\qquad +E\left\{\sum_{i=1}^{t}\left(h(Y_{\Psi_i})-x_{\Psi_i}\right)l_0(F_i(Y_{\Psi_1}^{\Psi_{i-1}}))|\mathcal{F}_t^{\Psi}\right\} \\
&= E\left\{\left(h(Y_{\Psi_{t+1}})-x_{\Psi_{t+1}}\right)|\mathcal{F}_t^{\Psi}\right\}l_0(F_{t+1}(Y_{\Psi_1}^{\Psi_t})) \\
&\qquad +\sum_{i=1}^{t}\left(h(Y_{\Psi_i})-x_{\Psi_i}\right)l_0(F_i(Y_{\Psi_1}^{\Psi_{i-1}})) \\
&= E\left(h(Y_{\Psi_{t+1}})-x_{\Psi_{t+1}}\right)l_0(F_{t+1}(Y_{\Psi_1}^{\Psi_t}))+m_t \\
&= m_t, \qquad (122)
\end{aligned}$$

where the third equality is since $Y_{\Psi_{t_0}}$ is $\mathcal{F}_t^{\Psi}$ measurable for any $t_0 \leq t$, the fourth is since $h(Y_{\Psi_{t+1}})-x_{\Psi_{t+1}}$ is independent of $\mathcal{F}_t^{\Psi}$ and the fifth is since $h(Y_{\Psi_{t+1}})$ is an unbiased estimate for $x_{\Psi_{t+1}}$. Hence, $(m_t, \mathcal{F}_t^{\Psi})$ is a zero-mean martingale (note that $Em_1 = 0$). Analogously, $\sum_{i=1}^{t}\left(x_{\Psi_i}-h(y_{\Psi_i})\right)l_1(F_i(y_{\Psi_1}^{\Psi_{i-1}}))$ is also a zero-mean martingale with respect to $\mathcal{F}_t^{\Psi}$, which completes the proof.

References

[1] B. Natarajan, K. Konstantinides, and C. Herley, "Occam filters for stochastic sources with application to digital images," *IEEE Trans. Signal Processing*, vol. 46, no. 5, pp. 1434–1438, May 1998.

[2] C.-H. Lamarque and F. Robert, "Image analysis using space-filling curves and 1D wavelet bases," *Pattern Recognition*, vol. 29, no. 8, pp. 1309–1322, 1996.

[3] A. Krzyzak, E. Rafajlowicz, and E. Skubalska-Rafajlowicz, "Clipped median and space filling curves in image filtering," *Nonlinear Analysis*, vol. 47, pp. 303–314, 2001.

[4] L. Velho and J. M. Gomes, "Digital halftoning with space filling curves," *Computer Graphics*, vol. 25, no. 4, pp. 81–90, July 1991.

[5] E. Skubalska-Rafajlowicz, "Pattern recognition algorithms based on space-filling curves and orthogonal expansions," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1915–1927, July 2001.

[6] T. Asano, D. Ranjan, T. Roos, E. Welzl, and P. Widmayer, "Space-filling curves and their use in the design of geometric data structures," *Theoretical Computer Science*, vol. 181, pp. 3–15, 1997.

[7] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve," *IEEE Trans. Knowledge and Data Engineering*, vol. 13, no. 1, pp. 124–141, January/February 2001.

[8] A. Bogomjakov and C. Gotsman, "Universal rendering sequences for transparent vertex caching of progressive meshes," *Computer Graphics Forum*, vol. 21, no. 2, pp. 137–148, 2002.

[9] R. Niedermeier, K. Reinhardt, and P. Sanders, "Towards optimal locality in mesh-indexings," *Discrete Applied Mathematics*, vol. 117, pp. 211–237, 2002.

[10] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 1, pp. 2–8, January 1986.

[11] N. Merhav and T. Weissman, "Scanning and prediction in multidimensional data arrays," *IEEE Trans. Inform. Theory*, vol. 49, no. 1, pp. 65–82, January 2003.

[12] A. Cohen, N. Merhav, and T. Weissman, "Scanning and sequential decision making for multi-dimensional data - part I: the noiseless case," to appear in *IEEE Trans. on Inform. Theory.*

[13] A. Cohen, *Topics in scanning of multidimensional data*, Ph.D. thesis, Technion, Israel Institute of Technology, 2007.

[14] T. Weissman, N. Merhav, and A. Somekh-Baruch, "Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1849–1866, July 2001.

[15] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2151–2173, September 2001.

[16] T. Weissman and N. Merhav, "Universal prediction of random binary sequences in a noisy environment," *Ann. App. Prob.*, vol. 14, no. 1, pp. 54–89, February 2004.

[17] T. E. Duncan, "On calculation of mutual information," *SIAM Journal of Applied Mathematics*, vol. 19, pp. 215–220, July 1970.

[18] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

[19] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1253–1264, April 2007.

[20] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: known channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 5–28, January 2005.

[21] E. Ordentlich and T. Weissman, "On the optimality of symbol-by-symbol filtering and denoising," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 19–40, January 2006.

[22] D. Guo, *Gaussian Channels: information, estimation and multiuser detection*, Ph.D. thesis, Princeton University, 2004.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[24] S. Verdú and T. Weissman, "The information lost in erasures," submitted to *IEEE Trans. on Inform. Theory*, 2007.

[25] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002.

[26] H. Helson and D Lowdenslager, "Prediction theory and Fourier series in several variables," *Acta Math.*, vol. 99, pp. 165–202, 1958.

[27] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 2nd edition, 1984.

[28] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267–271, April 1965.

[29] V. G. Vovk, "Aggregating strategies," *Proc. 3rd Annu. Workshop Computational Learning Theory,* San Mateo, CA, pp. 372–383, 1990.